Subclonal reconstruction of tumors by using machine learning and population genetics

Giulio Caravagna[1], Timon Heide[1], Marc J. Williams[2], Luis Zapata[1], Daniel Nichol[1], Ketevan Chkhaidze[1], William Cross[2], George D. Cresswell[1], Benjamin Werner[1], Ahmet Acar[1], Louis Chesler[3], Chris P. Barnes[4], Guido Sanguinetti[5,6], Trevor A. Graham[2]✉ and Andrea Sottoriva[1]✉

Most cancer genomic data are generated from bulk samples composed of mixtures of cancer subpopulations, as well as normal cells. Subclonal reconstruction methods based on machine learning aim to separate those subpopulations in a sample and infer their evolutionary history. However, current approaches are entirely data driven and agnostic to evolutionary theory. We demonstrate that systematic errors occur in the analysis if evolution is not accounted for, and this is exacerbated with multi-sampling of the same tumor. We present a novel approach for model-based tumor subclonal reconstruction, called MOBSTER, which combines machine learning with theoretical population genetics. Using public whole-genome sequencing data from 2,606 samples from different cohorts, new data and synthetic validation, we show that this method is more robust and accurate than current techniques in single-sample, multiregion and longitudinal data. This approach minimizes the confounding factors of nonevolutionary methods, thus leading to more accurate recovery of the evolutionary history of human cancers.

Cancers change over time through a process of clonal evolution[1], inevitably resulting in intratumor heterogeneity[2]. Genome sequencing of one or more bulk samples from tumors has become the most common way to study clonal evolution in human malignancies, and studies are dedicated to the identification of cancer (sub)clones[3]. A cancer 'clone' remains a loosely defined entity, and its purest definition is 'a group of cells within the tumor that share a common ancestor'. In phylogenetic terms, this would represent a monophyletic clade. However, this implies that any ancestor in the entire phylogenetic tree of a tumor can be identified as the founder of a distinct clone, even though it may show no biological difference from the rest of the cancer cells. This is why, in the field, we implicitly identify clones of interest, such as those that have growth/survival advantage (an ancestor under positive selection) or those that generate metastases (an ancestor that arrived and grew at a given metastatic site). The limits in the definition of a clone are important to bear in mind when attempting to recover the tumor clonal architecture.

To identify clones in bulk cancer samples, the established approach is unsupervised clustering of variant read counts[4], with each of the resulting clusters defined as a clone. This procedure, called subclonal reconstruction, leverages variant read counts and the associated variant allele frequency (VAF) of somatic mutations, adjusted for copy-number status and tumor purity, to identify groups of variants with similar cellular proportions. Subclonal reconstruction allows tracing of the life history of a tumor via determination of its phylogenetic tree (sometimes called a clone tree)[3].

Current methodologies approach subclonal reconstruction with sophisticated mixture models[4], implemented via Dirichlet processes[3,5,6] or Dirichlet finite mixtures[7]. These machine-learning methods are entirely data driven and are usually chosen because of their convenient statistical properties, rather than their adherence to the mechanisms of tumor evolution. They can be efficient and accurate, as long as the underlying assumptions are correct. All current subclonal reconstruction methods assume that variant read counts from bulk tumor samples present as a mixture of binomial or beta-binomial mutational clusters, each one corresponding to a clone. However, these are not the only observable patterns in the data: the mutations that occur within each clone while it expands are also detectable. Given the size of the human genome, even with low mutation rates (for example, substitutions of 10^{-9} nucleotides per base per division[8]), new mutations are expected at each cell division, and thus large numbers of passenger mutations inevitably accumulate within an expanding clone. The evolutionary dynamics of this passenger mutation accumulation are neutral, and give rise to a power-law-distributed 'tail' of ever more mutations at ever lower frequency. This has been mathematically demonstrated in theoretical population genetics[9–14] and is corroborated by genomic data at high resolution[15,16]. These within-clone neutral tails have not been directly addressed by previous methods, potentially confounding the measurement of clonal heterogeneity.

In the present study, we reconciled data-driven machine-learning approaches to clustering VAFs and corresponding cancer cell fractions (CCFs), with the insight given by evolutionary theory. Specifically, we combined Dirichlet mixture models with the distributions predicted by theoretical population genetics models[9–12], producing a model-based method for subclonal reconstruction called model-based clustering in cancer (MOBSTER). MOBSTER can process mutant allelic frequencies to identify and remove neutral tails from the input data, so that machine-learning subclonal

[1]Evolutionary Genomics and Modelling Lab, Centre for Evolution and Cancer, The Institute of Cancer Research, London, UK. [2]Evolution and Cancer Lab, Barts Cancer Institute, School of Medicine and Dentistry, Queen Mary University of London, London, UK. [3]Division of Clinical Studies, The Institute of Cancer Research, London, UK. [4]Department of Cell and Developmental Biology and UCL Genetics Institute, University College London, London, UK. [5]School of Informatics, University of Edinburgh, Edinburgh, UK. [6]International School for Advanced Studies, Trieste, Italy. ✉e-mail: t.graham@qmul.ac.uk; andrea.sottoriva@icr.ac.uk

reconstruction algorithms can be applied downstream to find subclones from read counts. We also expanded MOBSTER to analyze data from multiple samples of the same tumor, collected both in space and in time.

Results

Mutation, drift and selection. Cancers grow from a single cell, and hence neutral mutations that occur in the first few cell divisions are present at a high frequency in the final population, irrespective of the action of selection. In addition, stochastic fluctuations in population size of cell lineages can also increase the frequency of mutations in the absence of selection; this is called genetic drift[17]. The same is true within (sub)clones: a clone originates as a single cell, and neutral mutations that occur early within the clone are found in a large proportion of the clone's cells. Fundamental insight into the accumulation of mutations in the absence of positive selection came from the study of the Luria–Delbruck model in bacteria[18]. This has led to well-established population genetics theory describing the accumulation of mutations within neutrally growing populations[10,11]. The same theory applies to cancer clones[9,12] and can be extended to include positive selection[16]. Theory states that we should expect a tail of neutral passenger mutations within a clone (Fig. 1a). Neutral tails only recently became evident in cancer data with the adoption of high-depth whole-genome sequencing (WGS), as lower-depth sequencing (for example, <60×) is insufficient to detect tails reliably[16], and exome or panel sequencing often assay too few mutations to show a clear VAF spectrum.

Figure 1a shows the simplest example of a uniform 'neutral' tumor expansion. The corresponding clone tree has a single truncal node (Fig. 1b). The VAF spectrum for this tumor consists of a clonal peak at a high frequency, corresponding to the mutations that are present in all cells (that is, in the most recent common ancestor), and a neutral tail of mutations at lower VAFs generated as the clone expands (Fig. 1c). In the case where a subclone with selective advantage is present (Fig. 1d,e), the data will present as two peaks at high frequency (one clonal and one subclonal) as well as a mixture of two overlapping neutral tails[16] (Fig. 1f). Performing subclonal reconstruction on these data, assuming a generative mixture of just binomial or beta-binomial distributions, will detect several clusters within the neutral tail that are erroneously identified as subclones, as illustrated in two simulated cases (neutral in Fig. 1g and with one selected subclone in Fig. 1h). Importantly, mutations in neutral tails are not monophyletic, and hence grouping them together into clones is erroneous even under the strictest definition of a clone. Furthermore, when these incorrect clones are used downstream for phylogenetic reconstruction, the resulting trees (Fig. 1i) have a very different structure from the true trees (Fig. 1b,e), thus propagating errors and uncertainty in the tree construction, with many equivalent (but wrong) trees potentially fitting the same data.

Moreover, low-depth sequencing and low-purity data cause neutral tails to be undersampled and likely to be mistaken for subclones, because they lose their characteristic power-law shape. Simulated WGS data (Fig. 1j) show that, with low coverage or purity, the signal of a neutral tail becomes statistically difficult to distinguish from that of a selected subclonal cluster (Fig. 1k). This observation indicates that sequencing depth below 90×/100× and low purity prevents reliable subclonal reconstruction. We note that patterns of noisy subclonal VAF distributions that may represent undersampled tails (for example, Fig. 1k), are commonly observed in cancer-sequencing data at depth <90×/100×.

Model-based clustering of variant allele frequencies. The frequency f of newly acquired passenger mutations in an expanding population follows a Landau distribution[10], which at the frequency range detected by current sequencing standards can be approximated by a power-law distribution $X \approx 1/f^2$ (Fig. 2a), as we previously reported[9]. Subclonal alleles under positive selection, together with their hitchhiking passengers, will instead form clusters in the VAF distribution as they rise in frequency due to positive selection[16,19].

We can model VAFs or fraction data via beta distributions[7], and model read counts with binomial or beta-binomial distributions[3,5–7]. In MOBSTER (Fig. 2a), we model the evolutionary dynamics of a growing tumor containing subclones by combining beta distributions (expected from subclones under selection) with a power law (expected from neutral tails). After fitting the VAF distribution, tail mutations can be removed and clustering of read counts from the remaining mutations can be performed via standard methods (Fig. 2b). MOBSTER controls for tails while retaining the original variance of the data when clustering non-tail read counts downstream. Notably, MOBSTER always compares the fit of a mixture of clones with and without a neutral tail and uses a regularized model selection strategy to determine the best model fit to the data.

MOBSTER combines one Pareto type I random variable (a type of power law) with k beta random variables, resulting in a univariate finite mixture with $k+1$ components. The likelihood for n datapoints x_i is

$$p(D|\boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{i=1}^{n} \left[\pi_1 g(x_i|x_\star, \alpha) + \sum_{w=2}^{k} \pi_w h(x_i|a_{w-1}, b_{w-1}) \right],$$

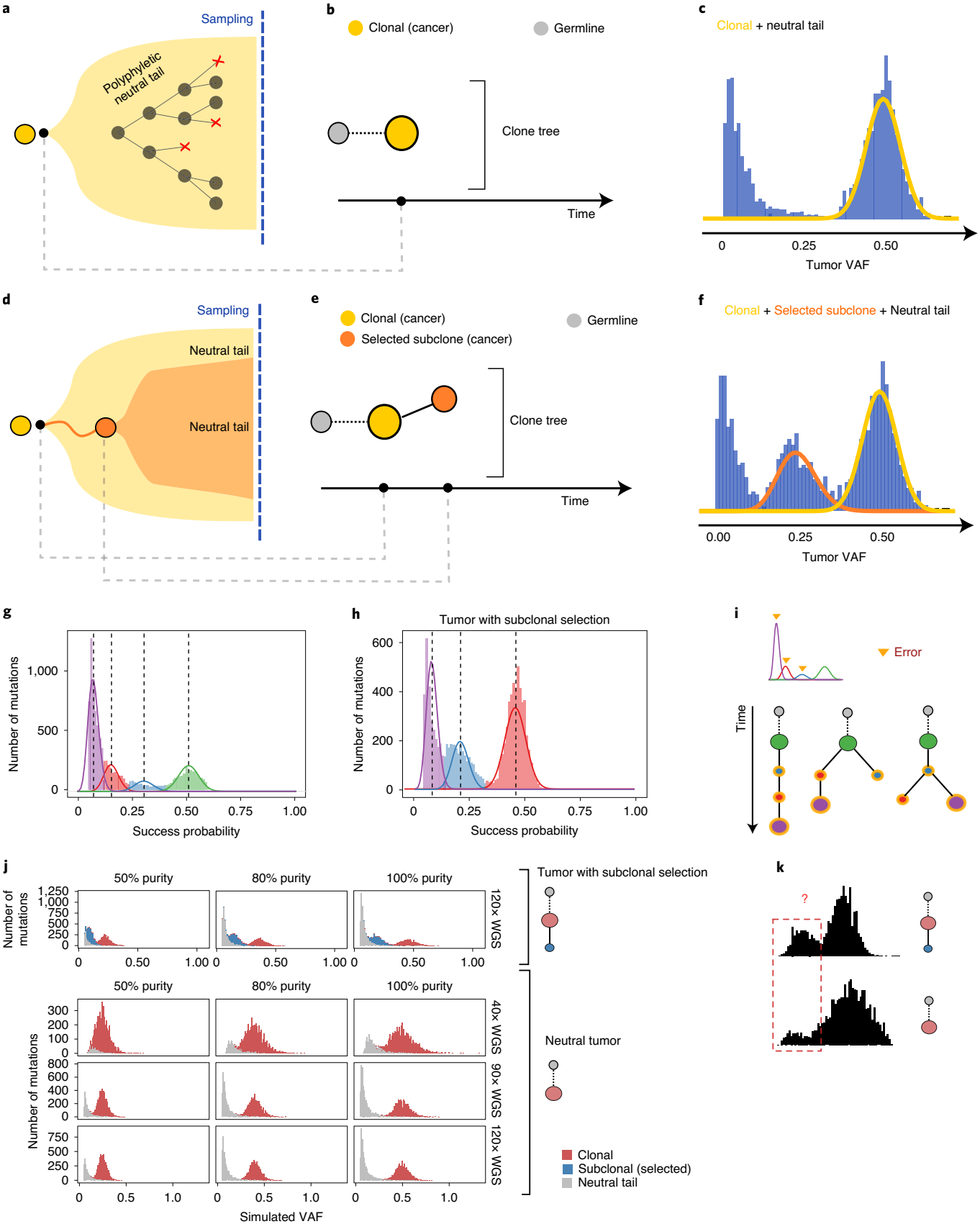
where g and h are density functions, $\boldsymbol{\theta} = (x_\star, \alpha, a_1, .., a_k, b_1, \ldots, b_k)$ is a vector of parameters and $\boldsymbol{\pi}$ is mixing proportions in a standard setting with $n \times (k+1)$ latent variables. The Pareto component follows $g(x|x_\star, \alpha) \propto 1/x^{1+\alpha}$ for $x \geq x_\star$, and the beta follows $h(x|a, b) \propto x^{a-1}(1-x)^{b-1}$ in [0,1]. A derivation of MOBSTER, its relation to other approaches and technical comments are available in the online content.

In the hypothetical example of a functionally monoclonal tumor with neutral subclonal dynamics (Fig. 1a), MOBSTER fits $k=1$ beta clusters of truncal mutations (present in all cancer cells) plus a neutral tail (Fig. 2c). Similarly, for a tumor with one selected subclone (Fig. 1d), MOBSTER fits $k=2$ beta clusters and a tail (Fig. 2d). When we identify and remove tail mutations from

Fig. 1 | Theoretical predictions of cancer genomic data under different evolutionary dynamics. a, A tumor formed by a single functionally monoclonal expansion follows neutral evolutionary dynamics driven only by mutation and drift. **b**, The clone tree can be represented as a single truncal clone. **c**, In diploid regions, the VAF distribution is characterized by one clonal cluster and a neutral $1/f^2$ tail of subclonal mutations. **d**, In a tumor with one subclone under positive selection (functionally polyclonal), the evolutionary forces of mutation and drift are still at play within each clone. **e**, The clone tree is represented as a truncal node giving rise to a selected subclone within it. **f**, The VAF shows one extra cluster due to subclonal mutations in the subclone that has risen in frequency due to selection. **g**, Standard subclonal deconvolution identifies clusters of neutral tail mutations that are not subclones, because they represent admixed polyphyletic lineages. In this panel, we show such clustering for the monoclonal tumor in **c**. **h**, Standard subclonal deconvolution for the tumor in **f**. **i**, This causes inflated estimates of the number of clones that propagate errors and uncertainty downstream, with several incorrect phylogenetic trees fitting the data. **j**, In these synthetic examples, the VAF distribution of a tumor with and without subclonal selection changes for different values of coverage and purity, affecting the ability to observe neutral tails. A neutral tail (gray) becomes difficult to detect at 40× depth. **k**, The degenerated tail at 40× can be statistically indistinguishable from a positively selected subclonal cluster. Data at such resolution are not powered to distinguish true positive subclonal selection from neutral tail mutations.

the data, subsequent clustering of read count mutations identifies the true tumor clones and their correct clone trees (inner clone tree panels).

Synthetic validation of the method and confounding factors. We used synthetic data to validate MOBSTER and quantify the degree to which neutral tails confound subclonal deconvolution



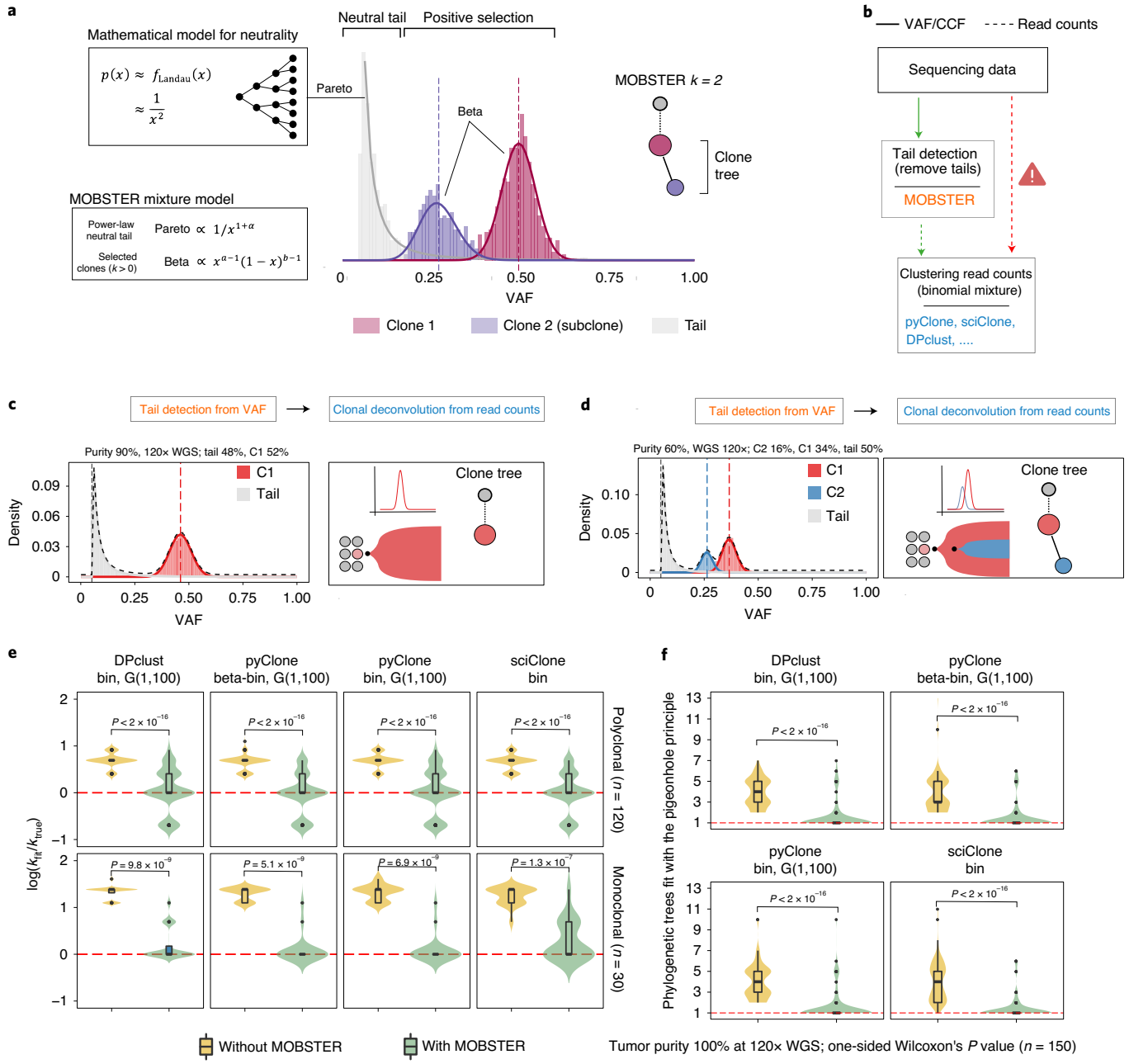


Fig. 2 | Model-based tumor subclonal reconstruction. a, MOBSTER combines a Pareto type I distribution with k beta random variables into a univariate finite mixture with $k+1$ components. The Pareto captures the frequency spectrum of neutral mutations predicted by theory (Landau distribution decaying as $1/f^2$), whereas beta components detect alleles under positive selection. The histogram shows clustering assignments for a tumor with one selected subclone ($k=2$). **b**, MOBSTER filters out neutral tail mutations, and one can cluster the rest with any tool for subclonal reconstruction using read counts. **c**,**d**, MOBSTER applied to the examples in Fig. 1a,b, respectively, detects the clusters corresponding to the true selected clones, hence recovering the correct clonal architecture. **e**,**f**, We used synthetic 120× WGS data from $n=150$ simulated tumors to compare current methods with MOBSTER (plots show mean and interquartile range (IQR), upper whisker is the third quartile +1.5×IQR and lower whisker is the first quartile −1.5×IQR). We measured how many clusters (**e**) and clone trees (**f**) we identify. Tests compare binomial mixtures from DPclust, pyClone and sciClone, and beta-binomial mixtures from pyClone, parameterized by concentration $\alpha > 0$. DPclust and pyClone learn α from the data assuming a gamma prior. sciClone is a variational method with hardcoded α. In **e** we report the logarithm of the ratio between the number of subclones found by MOBSTER (k_{fit}) and the true number of clones (k_{true}). The red dashed line represents $k_{\text{fit}} = k_{\text{true}}$. In **f** we plot the number of trees that can be fit by the pigeonhole principle using the output of each tool.

with standard methods (Supplementary Note and Supplementary Figs. 1–9). We used a stochastic branching process[16] to simulate the growth of $n=150$ tumors (online content and Supplementary Data). Out of these 150 cases, 30 tumors were neutral (as in Fig. 1a) and 120 contained one selected subclone (as in Fig. 1d). For each tumor we simulated bulk WGS at 120× median coverage and 100% purity. In every test, we always compared the fit of MOBSTER with and without a tail, retaining the best; we then recorded the predicted number of selected clones, k, and the fit precision (Supplementary Figs. 3 and 4). We note that, by applying further population genetics theory[16] to the output of MOBSTER, we can estimate the tumor evolutionary parameters, such as the mutation rate, the time of

emergence of subclones and their selection coefficients (Supplementary Fig. 5). We also carried out several other tests for the detection of low-frequency subclones admixed with tails (Supplementary Figs. 6 and 7).

By accounting for neutral tails, MOBSTER significantly outperformed standard approaches based on both Dirichlet variational mixtures and Dirichlet processes (Extended Data Fig. 1), two statistical frameworks at the core of subclonal reconstruction tools such as sciClone[7], pyClone[5], DPclust[3] and many others. Results are consistent for various parameterizations, in particular of the concentration parameter $\alpha > 0$, which determines the propensity of adding clusters to the fit[3]. In Fig. 2e we report the error rates for the inferred number of clones (k) with DPclust, pyClone (binomial and beta-binomial) and sciClone. The detection of spurious extra clusters caused high uncertainty around the clone tree, with many solutions fitting the data equally well (Fig. 2f). We tested the effects of sequencing coverage and purity on tail detection, and found that ~100× coverage and high purity were required to systematically identify tails. Higher coverage is required for samples with lower purity (Extended Data Fig. 1). Additional synthetic tests with complex clonal architectures confirmed the robustness of the method (Supplementary Figs. 8 and 9). These analyses indicate that the previously published moderate-depth WGS studies were underpowered to detect reliable subclonal architectures, because the signal used to distinguish a tail from a subclone deteriorates with lower sequencing depth (Fig. 1j). With adequate data and controlling for neutral tails, we found the correct number of clones in the large majority of tests. Not considering neutral tails led to a systematic pattern of errors that, in the worst cases, could lead to a fourfold overestimation of the number of clones.

Not accounting for neutral tails also significantly impacts multiregion sequencing, as we discuss in the Supplementary Note. We found that multiregion bulk sequencing is affected by confounders that originate from the spatial effects of tumor growth and spatial sampling bias. In multi-sample analyses (Supplementary Note) we characterized a confounder termed the 'hitchhiker's mirage' (Extended Data Fig. 2) caused by parts of neutral tails that spread in space, and that current methods mistake for selected subclones (Supplementary Fig. 10). We also characterized two additional confounders due to the presence of locally sampled ancestors (Extended Data Fig. 3) and admixing of multiple lineages (Extended Data Fig. 4). These spatial confounders affect virtually all tumors (Supplementary Figs. 11–13). Therefore, the joint use of MOBSTER and other heuristics is necessary to interpret subclonal deconvolution results from multiregion samples (Extended Data Fig. 5 and Supplementary Fig. 14).

Analysis of genomic data from human samples. We applied MOBSTER to high-coverage (>100×) WGS data available in the public domain (Supplementary Note). We first reanalyzed the breast cancer sample PD1420a sequenced at ~188× from Nik-Zainal et al.[3]. Compared with the original analysis, which found three subclones, MOBSTER fits two subclones ($k = 3$) and places a neutral tail for the lowest frequency cluster (Fig. 3a). The sciClone analysis of read counts for non-tail mutations confirmed $k = 3$ binomial clusters (two selected subclones). Both linear and branching phylogenies could be fitted to the output, with the branching tree matching the original analysis[3]. The cluster that MOBSTER fits to a tail appears in multiple positions of the tumor tree in the original paper after phasing[3]. This is consistent with our analysis, because the tail is polyphyletic, and hence composed of a mixture of descendants of the different clones. We measured the evolutionary parameters of this tumor from the fits, finding concordant estimates with our previous work[16]. The mutation rate was $\mu = 3.5 \times 10^{-7}$ mutations per base per tumor doubling, subclones emerged at $t = 5.5$ (smaller subclone) and $t = 10.4$ (larger subclone) doublings, and had selective coefficients of $s = 0.3$ and $s = 0.66$, respectively.

We reanalyzed the acute myeloid leukemia sample sequenced at 320× WGS by Griffith et al.[20]. MOBSTER identifies $k = 3$ clusters (two subclones) and a neutral tail (Fig. 3b). The two subclones were also detected by Griffith et al.[20], and were confirmed running sciClone after MOBSTER. However, MOBSTER simplified the clonal architecture by removing one spurious low-frequency subclone. This improves the interpretation of these data, possibly explaining why the tail was the only cluster without a clear subclonal driver mutation. The measured mutation rate was $\mu = 9.9 \times 10^{-10}$ per base per tumor doubling, subclones emerged at $t = 22$ and $t = 27$, and selection coefficients were $s = 1.3$ and $s = 3$, respectively.

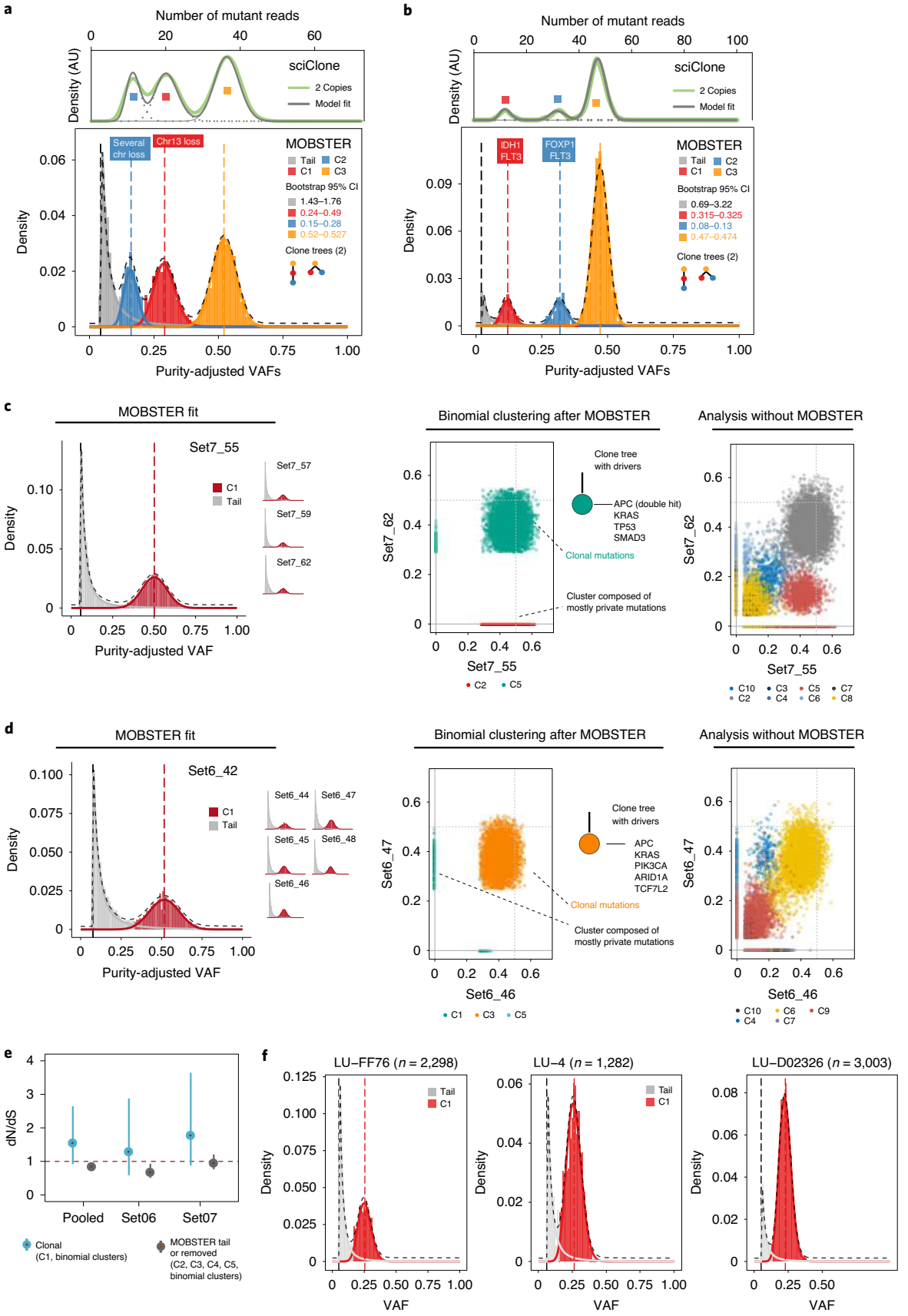
We also generated new multiregion WGS data (median 100×) from spatially separated regions of two primary colorectal cancers previously analyzed at lower depth in Cross et al.[21]. In tumor Set07 we analyzed high-confidence single nucleotide variants (SNVs) in diploid segments consistent across samples, and ran a comparative analysis with and without MOBSTER (Supplementary Note). The analysis with MOBSTER did not find evidence of positive subclonal selection (Fig. 3c and Supplementary Fig. 15), corroborated by the lack of subclonal drivers and truncal *APC*, *KRAS*, *SMAD3* and *TP53* mutations, as originally reported[21]. The analysis without MOBSTER would have depicted a complex subclonal structure, with several binomial clusters consistent with multiple clone trees (Fig. 3c and Supplementary Fig. 16). The analysis of Set06 gave similar results (Fig. 3d and Supplementary Fig. 17). In agreement with Cross et al.[21], the clone tree depicted a tumor with only truncal driver events in *APC*, *KRAS*, *PIK3CA*, *ARID1A* and *TCF7L2*, and neutral subclonal dynamics. Again, a standard analysis would have identified a complex clonal architecture with multiple subclones (Fig. 3d and Supplementary Fig. 18). Mutation rates were $\mu = 5.6 \times 10^{-7}$ for Set07 and $\mu = 4.3 \times 10^{-7}$ for Set06. Notably, orthogonal dN/dS analysis that uses the ratio of non-synonymous(dN) to synonymous (dS) mutations in order to detect selection[22,23] confirmed the lack of evidence for positive selection at the subclonal level in those tumors (Fig. 3e and Supplementary Note).

We also applied MOBSTER to $n = 3$ non-small-cell lung cancer samples sequenced at high depth (Fig. 3f). These three

Fig. 3 | Analysis of single-sample and multiregion whole-genome data. a, Breast carcinoma ~180× WGS sample from Nik-Zainal et al.[3]. MOBSTER identified a neutral tail plus $k = 3$ beta clusters (two subclones, consistent with two clone trees). Analysis of non-tail mutations with sciClone confirmed two subclones. The sciClone without MOBSTER would have fit one extra clone to the tail. A nonparametric bootstrap is used to estimate the 95% bootstrap confidence intervals (CIs) for the parameters. AU, arbitrary units. **b**, Leukemia ~320× WGS sample from Griffith et al.[20]. MOBSTER found two subclones ($k = 3$), confirmed with sciClone, and two clone trees. **c**, WGS data at 100× from four biopsies of colorectal cancer Set07. From the VAFs of diploid mutations we identified neutral tails and no subclonal selection; from non-tail mutations we found five clusters (multivariate clustering with $\alpha = 10^{-6}$, see Supplementary Note). C1 is the truncal cluster; all other clusters are enriched mutations private to a biopsy, indicating ancestor effect (Supplementary Note). The clone tree depicts a neutrally expanding tumor with all drivers in the trunk. Analysis without MOBSTER would have inflated the number of subclones (right panel; see Supplementary Figs. 20–23). **d**, WGS data at 100× from six biopsies of cancer Set06 also showed neutral subclonal dynamics. Without MOBSTER we would have inflated the number of selected subclones (right panel; see Supplementary Figs. 24–27). **e**, The dN/dS analysis for Set06 and Set07 comparing truncal versus subclonal mutations confirmed a lack of evidence for positive selection at the subclonal level, corroborating our conclusions. **f**, Three lung cancer cases from Lee et al.[24] sequenced at 100× WGS were consistent with neutral subclonal dynamics.

tumors were those with the highest coverage and purity among a recently published cohort[24] (see also low-purity cases in Supplementary Fig. 19).

Neutral evolution in 2,566 whole genomes from PCAWG. We reanalyzed, with MOBSTER, one of the largest available cohorts of cancer WGS data to date, collated by the Pan-Cancer Analysis of



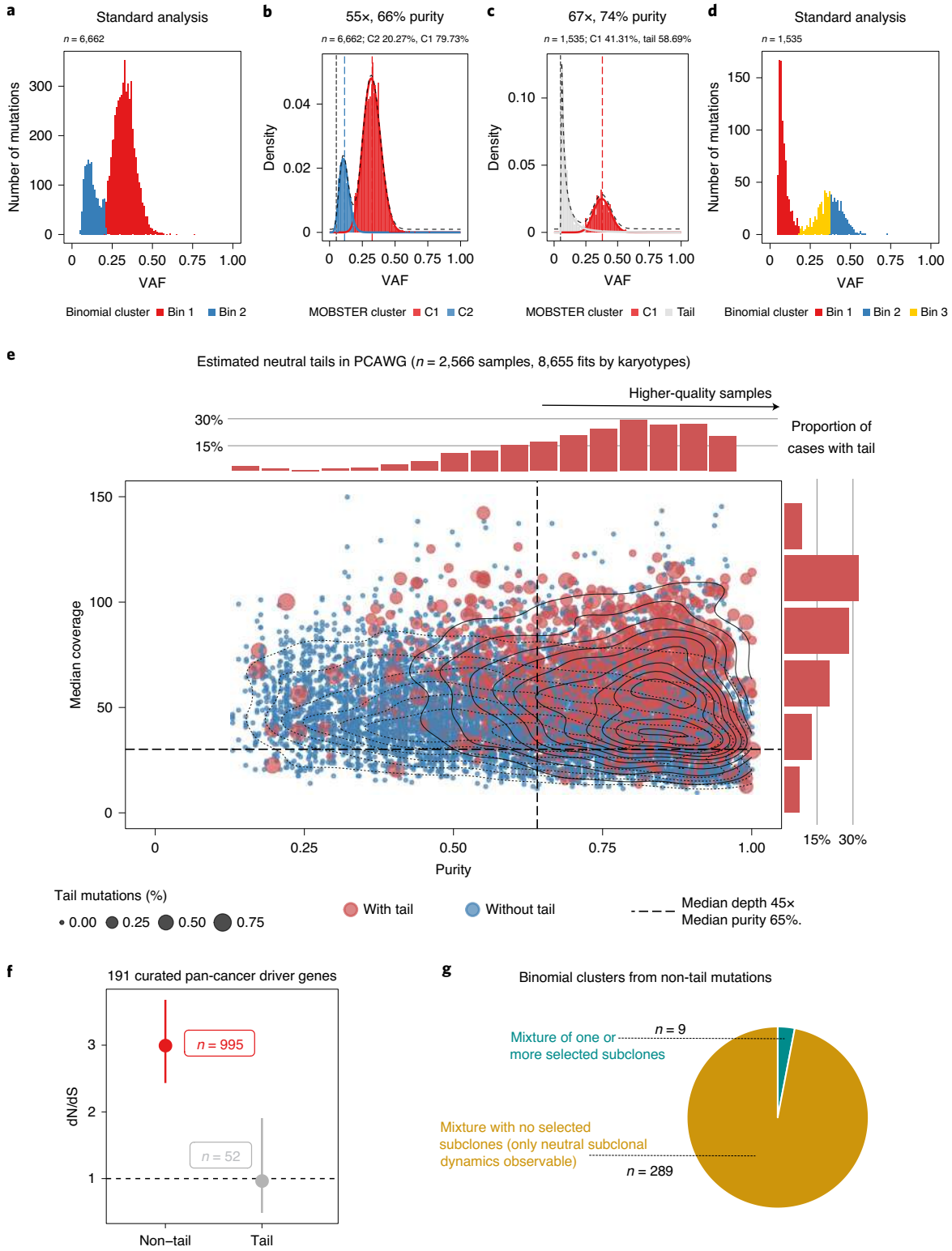


Fig. 4 | Analysis of 2,566 whole-genomes from PCAWG with MOBSTER. a, Fit of a PCAWG[25] tumor with 55× coverage and 66% purity using standard methods. **b**, At this data resolution, neutral tails are undersampled (Fig. 1j,k) and cannot be distinguished from selected subclones. **c**, In PCAWG cases with higher coverage (67×) and purity (74%), neutral tails can be clearly detected using MOBSTER. **d**, Analysis of the same tumor with standard methods would have identified multiple subclonal clusters, including a cluster of neutral tail mutations. **e**, We analyzed $n=2{,}566$ PCAWG samples, plotted here for purity versus coverage. Blue dots are tumors where MOBSTER cannot fit a tail; the red cases have a neutral tail. The percentage of tail mutations determines dot size and the marginal histograms report the normalized number of cases with a tail. **f**, We focused on the 902 diploid cases with coverage >30× and purity >65% (median of the cohort) where we could fit a tail. Using a panel of 191 pan-cancer driver genes, we show that tail mutations have $dN/dS \approx 1$, providing no evidence of positive selection (point estimate and confidence intervals from the dNdScv tool[22]). Clonal and subclonal non-tail mutations show $dN/dS > 1$, consistent with positive selection. **g**, If we take the 298 diploid cases with a tail containing at least 10% of the total mutational burden, we find evidence of a selected subclone only in nine cases (3% of tumors). Similar proportions are obtained if we impose a 5% or 2% cutoff on the size of the tail. See Supplementary Figs. 20–22.

Whole Genomes (PCAWG) international consortium and recently published in a series of studies[25], including the evolutionary history of more than 2,600 cancers[26]. The median depth of coverage in this dataset was 45×, with a median purity of 65%. According to our power analysis, data at this resolution are not suitable for reliable subclonal reconstruction (Fig. 1j,k and Extended Data Fig. 1). Figure 4a shows a PCAWG case in which a standard analysis called a selected subclone. The coverage was 55× and purity 66%, with a VAF distribution similar to the downsampled synthetic neutral cases shown in Fig. 1j. With these data, MOBSTER (Fig. 4b; more cases in Supplementary Fig. 20) cannot fit a neutral tail in the low-frequency portion of the VAF spectrum, and instead fits a subclone (Beta component). The ground truth is not known, but, given the resolution of the data, we cannot exclude the likelihood that subclonal mutations in this sample are the result of a degenerate neutral tail (Fig. 1j,k). In cases where coverage and purity were higher, MOBSTER did identify neutral tails and resolved the remaining clonal structure (Fig. 4c). As expected, standard approaches would have identified spurious clusters (Fig. 4d), thus compromising the whole subclonal reconstruction.

We found a widespread presence of neutral evolutionary patterns in PCAWG data using MOBSTER. We analyzed the VAF spectrum of 2,566 cancers (Supplementary Note). Theoretical population genetics predicts that, given enough power in the data, we should always expect to find a neutral tail, with or without selected subclones (Fig. 2a). However, we consistently found neutral tails only in samples with higher coverage and purity (Fig. 4e, red = cases with a neutral tail, blue = cases without a detectable tail), suggesting a lack of power for subclonal inference in most cases (Supplementary Fig. 21).

To further validate the presence of neutral tail mutations in this cohort, we focused on $n = 902$ near-diploid cancers with >30× depth, >65% purity and where a tail was detected. From these cases we identified somatic mutations mapping to putative cancer driver genes[25,26] in neutral tails versus non-tails and performed a dN/dS analysis[22] (Fig. 4f). This orthogonal measurement confirmed that mutations in tails were likely neutral ($dN/dS \approx 1$), aside from the caveats of interpreting dN/dS values in growing tumors[27], whereas non-tail mutations indicated selection ($dN/dS > 1$).

We then focused on $n = 298$ diploid cases that were found to have at least 10% of the total mutation burden in the tail, indicating sufficient power to detect the clonal architecture with confidence. We measured the proportion of tumors with a selected subclone, defined by two or more binomial clusters detected from non-tail mutations. We found evidence of ongoing subclonal selection only in $n = 9$ (3% of total; see Supplementary Fig. 22). In the remaining $n = 289$ cases, neutral evolutionary dynamics at the subclonal level were the adequate description of the data (Fig. 4g). Lowering the threshold for proportion of tail mutations did not change the results (5% tail = 2.7% non-neutral cases; 2% tail = 3.7% non-neutral cases).

Our analysis suggests that, for most PCAWG cases, the data resolution was too low to conduct robust subclonal reconstruction. Moreover, neutral tails were detectable in higher coverage and purity samples, indicating that neutral dynamics are often an adequate description of the observed subclonal heterogeneity. Standard analyses of these data therefore risk systematically mistaking neutral tails for subclonal clusters, thus inflating the complexity of the inferred subclonal architectures and producing incorrect phylogenetic trees. Our analysis using MOBSTER hence demonstrates that neutral evolutionary patterns are prevalent in PCAWG data.

Analysis of longitudinal whole-genome datasets. We analyzed a cohort of $n = 35$ matched primary-relapse glioblastoma samples from 16 patients profiled using ~100× WGS in a recent study by Körber et al.[28]. Our analysis identified nine cases characterized only by neutral evolutionary dynamics at the subclonal level in both primary and relapse, whereas seven patients had a detectable ongoing subclonal expansion (Supplementary Fig. 23). We found cases where positively selected subclones were unique to the primary or the relapse (Fig. 5a,b), but also cases where pre-existing subclones in the primary swept through the population in the relapse, probably due to positive selection from treatment (Fig. 5c,d). In some cases, we found evidence of multiple subclones at relapse (Fig. 5e,f). MOBSTER also identified clusters of mutations that were due to whole-genome duplications, as in the case of a diploid primary tumor that became tetraploid at relapse (Fig. 5g,h). We note that some of the confounding effects of neutral tails in multivariate analyses (Supplementary Note) were ubiquitous in these data and would have negatively impacted standard subclonal reconstruction (Supplementary Fig. 23). Orthogonal analysis with dN/dS (refs. [22,23]) methods suggested neutral values for tail mutations ($dN/dS \approx 1$) and positive selection for others ($dN/dS > 1$) using a panel of glioma driver genes (Fig. 5h). We note that the presence of subclones under positive selection in these data was also reported in the original study[28]. However, using MOBSTER we obtained simplified clonal architectures, pruning some of the clusters that were due to neutral tails. Indeed, a mixture of subclonal selection and neutral evolutionary dynamics through therapy has been recently reported in a large glioblastoma study[29].

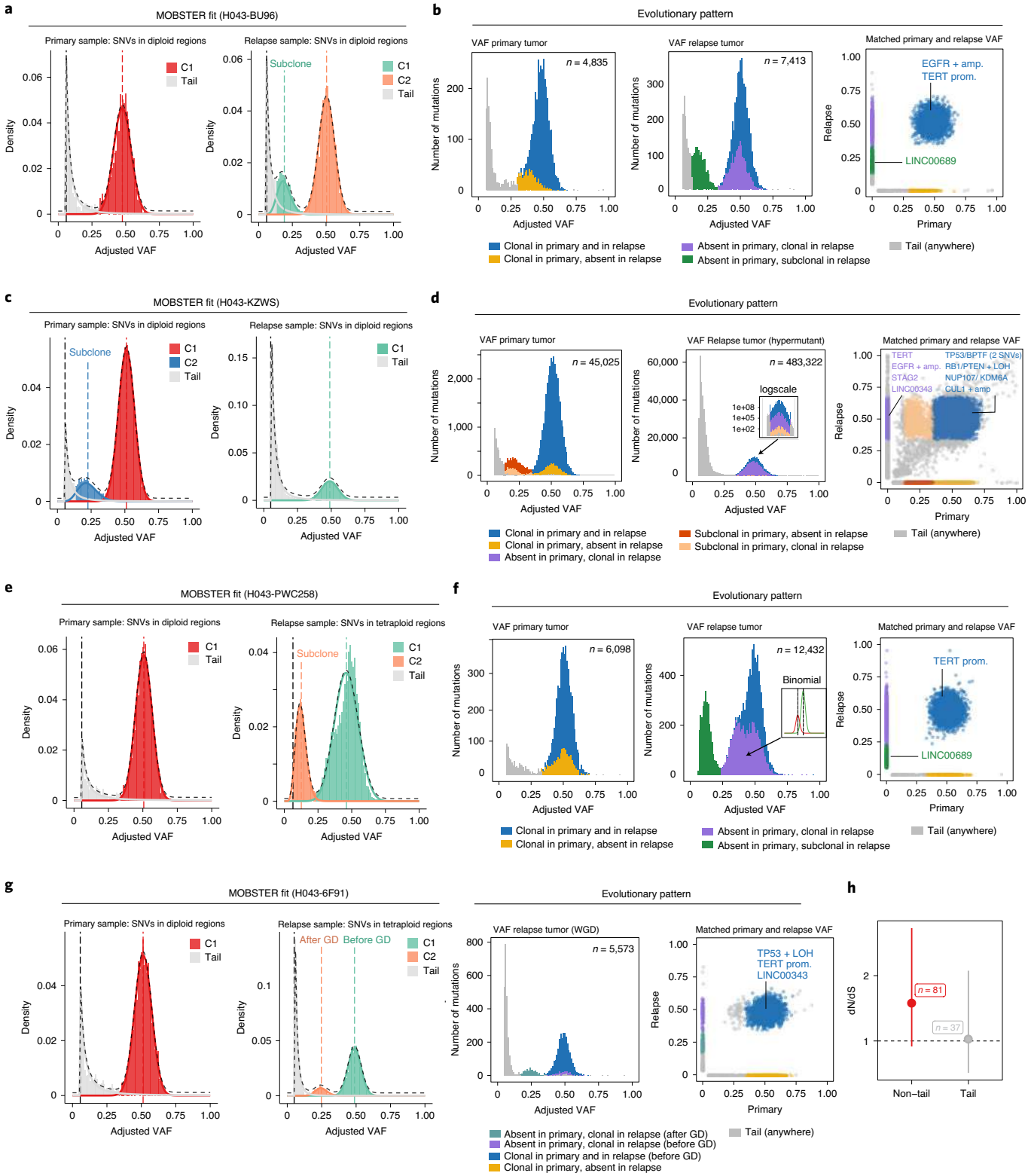
Discussion

Subclonal reconstruction from cancer bulk-sequencing data has paved the way for the study of cancer evolution[3,30]. Measurement of subclonal architectures also has clinical relevance: subclone multiplicity and other measures of intratumor heterogeneity have been reported as prognostic biomarkers[31–34]. Naturally, therefore, there is the need to ensure that subclonal reconstruction is accurate.

In the present study, we have presented a subclonal reconstruction method that combines data-driven machine learning with theoretical population genetics. This is in contrast to purely data-driven approaches that lack an underlying evolutionary model. Recently proposed standards for subclonal reconstruction[35] do not account for evolutionary dynamics, and hence this recommended best practice analysis is inherently flawed.

Fig. 5 | Analysis of longitudinal glioblastoma samples with MOBSTER. a, Patient H043-BU96 is one of $n = 16$ IDH-wild-type glioblastomas for which we analyzed WGS data (~100×) from pre-treatment and post-treatment longitudinal samples previously generated[28]. **b**, Analysis after MOBSTER identified subclones private to the primary (yellow) and relapse (green) tumors, respectively, the latter containing a putative driver mutation in LINC00689. **c**, Patient H043-KZWs MOBSTER fits. **d**, A subclone detected in the primary went on to sweep through the relapse, which was hypermutant after temozolomide treatment (zoom-in log(scale) panel). **e**, Patient H043–PWC258 MOBSTER fits. **f**, The primary sample showed neutral evolutionary dynamics, whereas the relapse contained detectable subclones possibly mixing with the neutral tail. An additional high-frequency subclone was detected from a downstream analysis using binomial clustering of read counts (purple cluster, split into two binomial components). **g**, MOBSTER can also be used to identify and assign clusters that are produced by whole-genome duplications, or more general aneuploid states. In such contexts, we expect to see peaks in the VAF distribution that distinguish mutations that happened before and after genome doubling (GD). In the case of patient H043-6F91, a diploid primary tumor (neutral) became whole-genome duplicated at relapse. **h**, Orthogonal dN/dS analysis (point estimate and confidence intervals from dNdScv) of mutations in 74 putative glioblastoma *GBM* driver genes assigned to neutral tails versus non-tails provided evidence of selection only in non-tail mutations. The full list of analyzed cases is available in Supplementary Fig. 23.

Moreover, we suggest that only high-depth sequencing data of >90/100× is appropriate to infer subclonal architectures, and even higher depth is required for purity <75%. Subclonal reconstruction from lower-depth data and lack of consideration for neutral tails risk a systematic overcalling of spurious subclones (Fig. 1j,k), leading to incorrect inference of the life history of tumors. These problems affect multiple previously published studies (for example, refs. [3,34,36]) and prohibit the inference of subclonal structures in the majority of PCAWG cases. Various issues arise also in multiregion-sequencing data, resulting from biases that are intrinsic to spatial sampling (Supplementary Note) and thus affect several previous studies that had insufficient depth of sequencing to infer metastatic spread (for example, refs. [37–39]). These issues also lead to inflated estimates of positive subclonal selection from VAF distributions. Single-cell sequencing removes the problem of admixing populations[40]; however, the underlying evolutionary dynamics



described by theory remains valid for the frequency of mutations among the N cells sequenced[41].

The major impact of MOBSTER is that it controls for neutrally evolving cancer cell subpopulations, cleaning up the signal for downstream analyses that seek to focus on functional intratumor heterogeneity. Given the wide use of clustering methods for subclonal reconstruction, MOBSTER has the potential to impact intratumor heterogeneity studies that use bulk sequencing, and even those that analyze the distribution of clade sizes in single-cell sequencing.

We also highlight the limitations of the definition of clone in cancer as a monophyletic clade with a most recent common ancestor, noting that, in the clinic, we are not interested in all the ancestors of a given group of cancer cells, but only in those few ancestors that drive progression, metastasis or treatment resistance. Importantly, even under this looser definition of a clone, clustering neutral tails with binomial models is incorrect and leads to the identification of false clones, mistaking the polyphyletic branching process that gives rise to neutral tails for a monophyletic lineage.

The present study highlights that there are intrinsic limitations to the information on tumor evolution encoded in current data, foremost being because of the systematic confounding factors caused by sampling complex three-dimensional tumors. We propose that our analysis represents a step toward a more refined approach to subclonal reconstruction in bulk cancer data, a necessity for genomic-aided precision medicine.

Accepted: 1 July 2020

References

1. Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481**, 306–313 (2012).
2. Turajlic, S., Sottoriva, A., Graham, T. & Swanton, C. Resolving genetic heterogeneity in cancer. *Nat. Rev. Genet.* **20**, 404–416 (2019).
3. Nik-Zainal, S. et al. The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
4. Dentro, S. C., Wedge, D. C. & Van Loo, P. Principles of reconstructing the subclonal architecture of cancers. *Cold Spring Harb. Perspect. Med*. **7**, a026625 (2017).
5. Roth, A. et al. PyClone: statistical inference of clonal population structure in cancer. *Nat. Meth*. **11**, 396–398 (2014).
6. Deshwar, A. G. et al. PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.* **16**, 35 (2015).
7. Miller, C. A. et al. SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Comput. Biol.* **10**, e1003665 (2014).
8. Lynch, M. et al. Genetic drift, selection and the evolution of the mutation rate. *Nat. Rev. Genet.* **17**, 704–714 (2016).
9. Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A. & Sottoriva, A. Identification of neutral tumor evolution across cancer types. *Nat. Genet.* **48**, 238–244 (2016).
10. Kessler, D. A. & Levine, H. Large population solution of the stochastic Luria–Delbruck evolution model. *Proc. Natl Acad. Sci. USA* **110**, 11682–11687 (2013).
11. Kessler, D. A. & Levine, H. Scaling solution in the large population limit of the general asymmetric stochastic Luria–Delbrück evolution process. *J. Stat. Phys.* **158**, 783–805 (2015).
12. Durrett, R. Population genetics of neutral mutations in exponentially growing cancer cell populations. *Ann. Appl. Probabil.* **23**, 230–250 (2013).
13. Nicholson, M. D. & Antal, T. Universal asymptotic clone size distribution for general population growth. *Bull. Math. Biol.* **78**, 2243–2276 (2016).
14. Griffiths, R. C. & Tavaré, S. The age of a mutation in a general coalescent. *Stoch. Models* **14**, 273–295 (1998).
15. Sun, R. et al. Between-region genetic divergence reflects the mode and tempo of tumor evolution. *Nat. Genet.* **49**, 1015–1024 (2017).
16. Williams, M. J. et al. Quantification of subclonal selection in cancer from bulk sequencing data. *Nat. Genet.* **50**, 895–903 (2018).
17. Hartl, D. L. & Clark, A. G. *Principles of Population Genetics* (Sinauer Associates, Inc., 2006).
18. Luria, S. E. & Delbrück, M. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* **28**, 491–511 (1943).
19. Graham, T. A. & Sottoriva, A. Measuring cancer evolution from the genome. *J. Pathol.* **241**, 183–191 (2017).
20. Griffith, M. et al. Optimizing cancer genome sequencing and analysis. *Cell Systems* **1**, 210–223 (2015).
21. Cross, W. et al. The evolutionary landscape of colorectal tumorigenesis. *Nat. Ecol. Evol.* **2**, 1661–1672 (2018).
22. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1–13 (2017).
23. Zapata, L. et al. Negative selection in tumor genome evolution acts on essential cellular functions and the immunopeptidome. *Genome Biol.* **19**, 924 (2018).
24. Lee, J. J.-K. et al. Tracing oncogene rearrangements in the mutational history of lung adenocarcinoma. *Cell* **177**, 1842–1857.e21 (2019).
25. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
26. Gerstung, M. et al. The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).
27. Williams, M. J. et al. Measuring the distribution of fitness effects in somatic evolution by combining clonal dynamics with dN/dS ratios. *eLife Sci.* **9**, 612 (2020).
28. Körber, V. et al. Evolutionary trajectories of IDHWT glioblastomas reveal a common path of early tumorigenesis instigated years ahead of initial diagnosis. *Cancer Cell* **35**, 692–704.e12 (2019).
29. Barthel, F. P. et al. Longitudinal molecular trajectories of diffuse glioma in adults. *Nature* **576**, 112–120 (2019).
30. Shah, S. P. et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486**, 395–399 (2012).
31. Andor, N. et al. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat. Med.* **22**, 105–113 (2016).
32. Morris, L. G. T. et al. Pan-cancer analysis of intratumor heterogeneity as a prognostic determinant of survival. *Oncotarget* **7**, 10051–10063 (2016).
33. Jamal-Hanjani, M. et al. Tracking the evolution of non-small-cell lung cancer. *N. Engl. J. Med.* **376**, 2109–2121 (2017).
34. Espiritu, S. M. G. et al. The evolutionary landscape of localized prostate cancers drives clinical aggression. *Cell* **173**, 1003–1013.e15 (2018).
35. Salcedo, A. et al. A community effort to create standards for evaluating tumor subclonal reconstruction. *Nat. Biotechnol.* **38**, 97–107 (2020).
36. Yang, L. et al. An enhanced genetic model of colorectal cancer progression history. *Genome Biol.* **20**, 168 (2019).
37. Yates, L. R. et al. Genomic evolution of breast cancer metastasis and relapse. *Cancer Cell* **32**, 169–184.e7 (2017).
38. Gundem, G. et al. The evolutionary history of lethal metastatic prostate cancer. *Nature* **520**, 353–357 (2015).
39. Noorani, A. et al. Genomic evidence supports a clonal diaspora model for metastases of esophageal adenocarcinoma. *Nat. Genet.* **347**, 1–10 (2020).
40. Navin, N. E. The first five years of single-cell cancer genomics and beyond. *Genome Res.* **25**, 1499–1507 (2015).
41. Chkhaidze, K. et al. Spatially constrained tumour growth affects the patterns of clonal selection and neutral drift in cancer genomic data. *PLoS Comput. Biol.* **15**, e1007243 (2019).

Methods

Model-based clustering of cancer subclones with MOBSTER. The subclonal deconvolution problem is popular in the cancer literature[35]. Given read counts for a list of mutations detected from bulk sequencing of multiple tumor samples, we want to detect clusters of mutations that represent cancer subpopulations admixed in our samples. The problem can be framed to include any type of somatic mutation for which we can estimate the frequency, in the data, of the somatic (that is, alternative) allele. Usually, the mutations that are easier to call are SNVs; more complex structural variations or insertion–deletions are more challenging to determine accurate allelic frequencies. Regardless of mutation types, our aim is to determine mutation clusters that suggest cancer subpopulations (that is, clones) under positive selection.

MOBSTER is a mixed method that combines two types of random variables to approach this problem.

The frequency spectrum and the observational process. Kessler and Levine[10] have shown that, in the large population solution of the stochastic Luria–Delbrück model, the probability of having m mutants follows a fat-tail Landau distribution:

$$p(m) = \frac{1}{\mu N} f_{\text{Landau}}\left(\frac{m}{\mu N} - \log \mu N + \gamma - 1\right).$$

Here N is population size, μ the mutation rate and γ the Euler constant. The asymptotic behavior of f_{Landau} can be approximated as $f_{\text{Landau}}(x) = 1/x^2$, which leads to the power-law approximation that has also been derived by others[12–14] as $p(m) \approx 1/m^2$.

A generative model for this power law can be constructed with a standard, Markovian stochastic, birth–death process of cell division—sometimes called a branching process[16]. The existence of patterns of neutral evolution is thus a consolidated result from population genetics arguments that describe the spread of alleles in growing populations without recombination, such as cancer[17]. In other words, the progeny of each clone accumulates neutral passenger mutations until any of their daughter cells acquires a new mutation that undergoes selection, because it triggers a new clonal expansion with increased fitness; the power-law spectrum therefore emerges by the frequencies of passengers. When a daughter cell enjoys a clonal expansion, however, the frequency of the variant alleles that accrued from the ancestor cell, to the actual cell that acquired the driver, will grow. Eventually, this new subclonal expansion will become detectable if selection forces are strong compared with the background (which is the clone within which this cell was born). In a recursive fashion, the progeny of this new cell/subclone will start dividing, giving rise to another power-law-distributed tail of within-clone neutral dynamics. Example subclonal evolutionary dynamics are shown in the Supplementary Data, where we animate a subclonal expansion, which shows how subclones emerge from low frequency up until they sweep, and how the allele frequency distribution changes over time.

Importantly, we want to make it clear that the power-law part of the spectrum—that is, the tail—results from the accumulation of passenger mutations in the progeny of each clone. We note that this result—in particular exponent 2 (shape)—refers to the total population structure of the tumor, which is accessible only in the theoretical scenario in which we can sequence all the cancer cells. Therefore, any specific finite sample that we collect and sequence, which is also contaminated by normal cells, might exhibit deviations from this theoretical distribution[16]. Deviations from strict exponential growth, for example, due to spatial constrains, can also cause theoretical deviations from exponent 2 (refs. [13,42]). However, we use this result to create a parametric model-based approach to analyze cancer data (that is, we fix the type of distribution, but not its parameters).

Input data and conceptualization. We work with sequencing data for the variant alleles of n somatic mutations, which we can pre-process in different ways. One option is to adjust VAF values for copy number and purity, retrieving the so-called CCFs and re-scaling them into [0,1] by halving the CCFs. With these adjusted VAF values we expect a clonal peak at roughly 50% VAF, with outliers spreading around 0.5 but well below 1; compared with CCFs, these values avoid the truncation of values >1 (ref. [3]). Another similar option is to adjust VAF values only by copy number, obtaining the so-called cellular prevalence. A third option is to use the raw VAF data directly; in this last scenario we can further split mutations by karyotype, that is, the absolute copy-number segments to which they map, and account for the fact that different aneuploidy states have different expected distributions (for example, a triploid tumor is expected to have two peaks of mutations, plus a tail and possibly subclonal clusters).

On real data, we suggest using mutations that map to copy-number segments with common karyotypes (that is, copy states), such as diploid regions (with or without loss of heterozygosity), and triploid and tetraploid segments. Mutations mapping to more complex karyotypes (for example, highly amplified oncogenes) can always be mapped post-hoc, after clustering, and should account for a small subset of the tumor's mutational burden. We stress the use of mutations in high-confident copy-number regions to carry out subclonal deconvolution; miscalled copy-number states confound the inference, creating artifact clusters of mutations. As best practice, we usually attempt a first fit using diploid genomes without losses of heterozygosity (that is, regions with one copy of the major and minor alleles), where we can identify high-confidence diploid SNVs.

Regardless of the representations, a model for the frequency spectrum ρ of the observed mutations with $k \geq 1$ detectable clones is a random variable that follows:

$$\rho \approx \sum_{i=1}^{k} (Y_i + B_i),$$

where

- $Y_i \propto x^{-\alpha}$ is a power-law random variable for frequencies of neutral mutations in the progeny of clone i. The generic exponent $\alpha > 0$ gives flexibility to accommodate all the confounders described above.
- $B_i \in [0,1]$ is a beta random variable modeling the signal of clone i. In layman's terms, B_i models the 'peak' in the VAF distribution due to the hitchhikers of the clone. These distributions range in [0,1], rendering them suitable to describe allelic frequencies (and also the motivation behind why we scale CCF values to fit this range). For the sake of simplification, we assume here working with adjusted VAF values, so that aneuploidy states (amplified, unamplified) are adjusted to form a single peak in the distribution (that is, exactly as with CCFs).

This model looks simple, and further observations are required to turn it into a mixture of standard random variables. In this formulation, the random variables for the tail and the bump of a clone are coupled to capture a joint signal. Although the overall mixing proportions can be assumed to be independent, this compound random variable requires an extra level of mixing within each clone, that is, another mixing weight to properly capture the proportions of the clone tail and bump. We can, however, simplify this model by accepting that tracking finer details is only for the clusters of each clone, which we use to identify subpopulations in the frequency spectrum (that is, we use the clone's peak, obtained from the cluster's mean, to assess the phylogenetic history of the tumor).

We therefore simplify the model by noting that all tails have the same exponent $\alpha > 0$, which holds if all clones have the same mutation rate. If the mutation rate does not change among subclones, that is, when there are no hypermutant subclones, all tails are described by the same theoretical distribution, and can be represented as multiple instances of the same random variable. Thus, we group them together in a single power-law tail:

$$\rho \approx \left(Y + \sum_{i=1}^{k} B_i\right).$$

Here the random variables have the same meaning as above, but the clone is no longer indexed by i. This model has a key advantage over the one in which each clone 'emits' its own tail: the random variables are decoupled and allow a simple mixture-model formulation, which we present in 'Distributions and likelihood'.

Before concluding, we observe that given that ρ, the observational model for read counts collected from next-generation sequencing, is a standard binomial process $n|\rho,m \approx \text{Bin}(n|m,\rho)$, where m is the coverage (total number of reads) and n the number of reads harboring the variant allele, then ρ is the success probability for m identical and independently distributed (i.i.d.) Bernoulli trials. It is important to observe that the frequency spectrum and observational process look at the data from different perspectives: the former is a distribution on allelic frequencies, whereas the latter is on read counts. In this observational model we can in principle use beta-binomial distributions to account for coverage overdispersion.

Relation to other models in the literature. The literature is rich with models that describe the above observational process and variations thereof, with either binomial distributions or beta-binomial distributions. We briefly discuss those that are more related to our framework.

Bayesian methods that employ Dirichlet processes for infinite binomial mixture models are a popular generalization of the observational process. These nonparametric methods can fit an unspecified number of clusters k to data, simplifying the model selection procedures. PyClone[5], DPclust[3] and PhyloWGS[6] are three popular tools for clonal deconvolution that, in different ways, use this framework: pyClone and DPclust implement binomial mixtures, with the former also supporting beta-binomial distributions; in both cases a stick-breaking construction for Dirichlet process priors is adopted[43]. PhyloWGS, instead, combines binomial distributions with a tree stick-breaking construction for the Dirichlet process priors[44], which allows PhyloWGS to cluster jointly the input SNVs, and construct a phylogenetic tree for the detected clones.

An alternative popular approach based on finite mixture models is sciClone[7], which supports binomial, beta and Gaussian mixtures. SciClone fits the models to data via variational inference, an information-theoretical approach to approximate the posterior distribution over the model's parameters. SciClone is a hybrid tool, because it can cluster allelic frequencies via beta/Gaussian mixtures, and read counts via binomial mixtures. We want to note that, with beta distributions, canonical Bayesian modeling leads to intractable priors, even if the conjugate prior distribution of the beta distribution can be found by following the principles of conjugate priors for the exponential family. For this reason, variational inference

of beta mixtures exploits a gamma approximation to the prior and posterior distributions, originally derived by Mao and Li[45]. In this approximation we cannot derive the so-called evidence lower bound, a standard measure to monitor convergence of a variational fitting algorithm.

These models are related to MOBSTER's framework: they assume that ρ can be approximated by a point process (for example, a Dirac distribution) centered at the beta means. The potential pitfall is clear: by applying the observational process to neutral mutations, the number of clones is overestimated. Clusters will be called from tail mutations (polyphyletic lineages), which is wrong when we look for clones under selection. We note that sciClone with beta distributions models the allele frequency spectrum as well; however, they do not account for power-law tails of neutrally evolving mutations.

Distributions and likelihood. MOBSTER implements a statistical model to fit n VAF values to Y, the tail, and to any one of the B_i betas, the clones (predefined in number). From a fit, tail mutations can be removed by inspecting clustering assignments, and other methods can be used to fit the observational process on the read counts of the remaining data. For this reason, MOBSTER is complementary to the tools mentioned above, because it works upstream of the observational process. Nevertheless, our method also provides a preliminary indication on the possible number of subclones in the tumor: with high-quality data with low dispersions, one can expect the same number of clones to be confirmed by downstream analysis of non-tail mutations.

The fit uses a pre-specified number of $k+1$ components, where Y is a Pareto type I distribution as the power-law tail. For a scale x_* and shape $\alpha > 0$, its density is

$$g(x|x_*, \alpha) = \alpha x_*^{\alpha} \frac{1}{x^{\alpha+1}}$$

for $x > x_*$, and 0 otherwise. Notice that the density is 0 for values below the scale parameter, which requires a sharp cutoff on the input VAF, and that its support is $[0,+\infty]$. The model also uses k beta distributions B_1, B_k to model clonal and subclonal clusters. For a shape $a > 0$ and $b > 0$ the density of a beta random variable is

$$h(x|a, b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a, b)}$$

where $B(a, b) = \int_0^1 x^{a-1}(1-x)^{b-1}\mathrm{d}x$ is the Beta function. The support of this distribution is $[0,1]$, the full frequency spectrum.

The overall model uses a Dirichlet prior on the abundance of each clone; thus MOBSTER is a finite Dirichlet mixture model with both beta and Pareto distributions. The model likelihood for a dataset $X = \{x_i | i = 1, \ldots, n\}$, where we assume each x_i to be i.i.d., is a combination of two types of densities:

$$p(D|\boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{i=1}^{n} \left[\pi_1 g(x_i|x_*, \alpha) + \sum_{w=2}^{k} \pi_w h(x_i|a_{w-1}, b_{w-1}) \right].$$

We use $\boldsymbol{\theta}$ as a shorthand to the model parameters, and $\boldsymbol{\pi} = [\pi_1 \ldots \pi_{k+1}]$ for the mixing proportions—a standard Dirichlet variable on the $(k+1)$-dimensional probability simplex. Notice that, just for notational convenience, we are assuming that the first model component is the Pareto random variable (the tail); we hold this setup fixed even if the model does not fit a tail (in that case we force $\pi_1 = 0$). As a result of this, we use the index $w-1$ for the parameters of the beta distributions just to reflect that their index starts from 1.

Fitting MOBSTER. The formulation uses $n \times (k+1)$ latent variables $\boldsymbol{z}$.

We fit the model parameters via a maximum likelihood estimation (MLE) through an adaptation of a standard expectation–maximization (EM) approach. This alternative is faster than a Bayesian Monte Carlo strategy, at the drawback of inferring a point estimate of the parameters. The lack of an explicit measure of uncertainty in the prediction (confidence) can be mitigated using the bootstrap.

We perform these steps to fit a MOBSTER model. In the E-step, we compute the posterior estimates of the latent variables as usual, once we account for the two different distributions involved:

$$z_{w,1}|\boldsymbol{\theta} \propto \pi_1 g(x_i|x_*, \alpha), \qquad z_{w,i}|\boldsymbol{\theta} \propto \pi_i h(x_w|a_i, b_i).$$

In both cases the normalization constant, C_w, is the overall density mass for point x_w:

$$C_w = \pi_1 g(x_w|x_*, \alpha) + \sum_{i=2}^{k} \pi_i h(x_w|a_{i-1}, b_{i-1}).$$

In the M-step, for the Pareto tail, we begin by noting that the scale x_* of the distribution can be set to its MLE[46], which is known to be the smallest observed frequency $x_* = \min X$. This is a constant of the data, so we have one less parameter to fit. We fit the Pareto shape α, given x_*; switching to the log(likelihood) and including latent variables, its MLE estimator is

$$\alpha_{\mathrm{MLE}} = -\frac{\sum_{i=1}^{n} z_{i,1}}{\sum_{i=1}^{n} z_{i,1} \log(x_*/x_i)}.$$

For the beta clones, in the M-step, the MLE estimator for the distributions has no closed form; we can resort to approximate it numerically, increasing the computational burden. We can also rely on a recent analytical result on the moment-matching (MM) estimator of mixtures of beta values by Schröder and Rahmann[47]. MM consists of matching t empirical moments of the data X to the theoretical moments of the distribution, and solving for them. Here $t = 2$ (mean and variance); a beta distribution has a mean μ and variance σ given by

$$\mu = \frac{a}{a+b}, \qquad \sigma = \frac{ab}{(a+b)^2(1+a+b)}.$$

For a beta, conditioned on the latent variables, the MM estimator is

$$\mu_{i_{\mathrm{MM}}} = \frac{\sum_{w=1}^{n} z_{w,i} x_w}{n\pi_i}, \qquad \sigma_{i_{\mathrm{MM}}} = \frac{\sum_{w=1}^{n} z_{w,i}(x_w - \mu)^2}{n\pi_i}.$$

Given estimates for μ_i and σ_i, we can re-parameterize the beta as

$$a_{i_{\mathrm{MM}}} = \left(\frac{1-\mu_i}{\sigma_i} - \mu_i^{-1} \right) \mu_i^2, \qquad b_{i_{\mathrm{MM}}} = \mu_i(\mu_i^{-1} - 1).$$

We remark that MM is not the same as computing the MLE, which computes the zeroes of the derivative of the likelihood with respect to the parameters $\boldsymbol{\theta}$, $\partial h/\partial \boldsymbol{\theta}$. Thus, the properties of a standard EM approach do not hold when we compute updates via MM: we cannot guarantee that the likelihood increases monotonically, because we cannot employ Jensen's inequality. It has, however, been shown[47] that the differences between the estimators are negligible in most cases. For the sake of precision, Schröder and Rahmann propose calling a fit through the MM for beta distributions the 'iterative method of moments', rather than EM.

In MOBSTER's implementation we provide both a standard EM fit with numerical solution for the MLE of beta distributions, and the faster iterative method of moments. In the former case, we monitor convergence of the likelihood, as standard. In the latter, we use the posterior estimates of $\boldsymbol{\pi}$ because the likelihood is not monotonically increasing. A theoretical property of this MM approach is that, in each step, before updating the component weights, the expectation of the estimated density equates the sample mean. In particular, this is true at a stationary point; a proof of this is in Lemma 1 of Schröder and Rahmann[47].

Initial conditions. As is standard in EM approaches, we compute the fit with several random initial conditions. We provide two heuristics to compute the initial condition of the fit (Supplementary Fig. 1). One is based on a peak detection heuristic applied in the frequency range [0.1,1] to VAF values binned with size 0.01. To detect k initial peaks we perform k-means clustering of each peak's x coordinate, and store their centers. If there are $w < k$ peaks to cluster, we sample $k - w$ random values in [0,1] for the remaining peaks. We use the centers of these clusters as the mean of k beta distributions with randomized variance sampled in $[10^{-3}, 0.25]$; we do sample variance values until the corresponding beta parameters a and b are positive. For the tail, α is randomly sampled in the interval [0.01,5]. These values provide wide ranges of different initial distributions. An alternative method for selecting the initial condition of the fit is totally randomized.

Experimental results show that peak detection is a more robust initialization method; the random counterpart sometimes leads to beta distributions with a mean approaching 1, a region of parameter values in which the likelihood becomes less stable, leading to numerical difficulties. In many cases, we test fits with both initial conditions and retain the best one.

Clustering assignments and model selection. We do not want the fit to be biased toward tails, because we would miss low-frequency subclones that hide in the tail. Besides, simulations suggest limits to the detectability of tails, and therefore we shall not assume the tail to be always present in the data. For this reason, MOBSTER can turn off the Pareto component of the mixture (that is, setting $\pi_1 = 0$) and fit just k beta. Hence, we can perform model selection for $1 \leq k \leq K$ considering both models with and without a tail. This induces a statistical competition and allows us to select the model that best explains the data, with or without a tail.

In MOBSTER we compute the negative log-likelihood $\mathrm{NLL} = -\log f(X|\boldsymbol{\theta}, \boldsymbol{\pi})$ of the data, which we use to derive the usual Akaike information criterion (AIC) and Bayesian information criterion (BIC) scores: $\mathrm{BIC} = 2\mathrm{NLL} + |\boldsymbol{\theta}| \log \boldsymbol{n}$ and $\mathrm{AIC} = 2\mathrm{NLL} + 2|\boldsymbol{\theta}|$.

These criteria favor simpler fits by penalizing a model for the number of its parameters $|\boldsymbol{\theta}|$. A model with k beta distributions and one tail has $|\boldsymbol{\theta}| = 3k + 2$ parameters ($k+1$ for the Dirichlet mixture $\boldsymbol{\pi}$, $2k$ for the beta(s) and 1 for the Pareto tail). The fit without tail model has $|\boldsymbol{\theta}| = 3k - 1$ parameters; fewer parameters reduce the penalty less, thus favoring fits without a tail.

In MOBSTER we want to drive the fit to select separate clusters, that is, fits with few overlapping components, which we do not achieve using BIC or AIC. We achieve these separations by using instead two types of entropy terms. In one case we compute, from the latent variables, the usual entropy $H(\boldsymbol{z})$:

$$H(\boldsymbol{z}) = \sum_{i=1}^{k+1} \sum_{j=1}^{n} z_{j,i} \log z_{j,i}$$

and obtain the standard integrative classification likelihood (ICL), ICL = BIC + $H(z)$, approximated through the BIC[48]. In this article we also introduce a heuristic variation to the ICL, which we call reICL, a reduced-entropy criterion in which we use the entropy of mutations that are not assigned to a tail (Supplementary Fig. 1). This is defined as reICL = BIC + H($\hat{z}$), where $\hat{z}$ is the latent variables for the set of mutations $\{x|1{\neq}\text{argmax } z_x\}$, re-normalized. Notice that, in practice, $\hat{z}$ is defined from the hard clustering assignments that we use to assign mutations to clusters; cluster '1' is the label to identify tail mutations.

Entropy terms in ICL and reICL help to fit separate clusters because overlapping mixture components have higher entropy, and therefore penalty. The maximum entropy distribution is the uniform one, which is when we cannot confidently assign mutations to clusters (a point that seems to be equally well explained by multiple components). By definition, ICL will push toward fits with a clear separation among tail and beta components, whereas reICL will require only separation of the beta ones. This modification to the ICL seems reasonable because the Pareto tail overlaps, by definition, all subclonal clusters, and this leads to strong entropy penalizations with ICL. For this reason, ICL will be more stringent in calling tails than reICL, which drops a part of the entropy penalty, restricting its computation to $\hat{z}$. See also Supplementary Fig. 1 for a graphical explanation.

Notice that, as we are using NLL, we seek to minimize these scores. In the tests, we investigate different model-selection strategies, and choose, as default score for model selection in MOBSTER, reICL, which seems to provide a nice tradeoff between the ability to identify the beta components and retention of the tail structure.

Analysis of synthetic data. In the Supplementary Note and Supplementary Data, we explain how we used branching processes to generate tumors without and with space, and present output metrics to assess precision and sensitivity of our analyses (number of clusters, confidence in the predictions, rates of false/true positives/negatives, the effect of coverage and purity, and the ability to identify subclones). In the tests we used MOBSTER and other tools for subclonal deconvolution.

We found MOBSTER and the analyses built around it to be accurate, across all simulated tumors. In all cases tails improve fit quality, from a statistical point of view. This clustering problem is challenging because tails and clones overlap, confounding weak signals of subclonal selection at the low-frequency VAF. We used our performance and combinations of coverage and purity to identify minimum requirements for reliable deconvolution in nonspatial data. In general, we assessed that we can fit subclones and tails for a wide range of parameter values, but overlapping distributions complicate the inference. MOBSTER does not show biases and can identify subclones, even when they have low VAFs (Supplementary Note).

From multiregion data (Supplementary Note) of polyclonal tumors, we identified three confounders that inflate the number of clones reported by a 'standard' analysis. The confounders contribute binomial clusters that cannot be directly linked to clonal evolution patterns originating from positive selection. Branching structures originating from the confounders are also misleading, and do not reflect selection-driven branched evolution. One of the confounders can be solved by MOBSTER; two require extra heuristics discussed in the Supplementary Note.

Analysis of patient-derived data. The description of all the data analyzed is in the Supplementary Note, as well as the Supplementary Data. All summary statistics for all fit samples of this article are available in Supplementary Table 1.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Data in Fig. 3a were from Nik-Zainal et al.[3]. Data in Fig. 3b were from Griffith et al.[20]. Data in Fig. 3c–e were cases from Cross et al.[21], here re-sequenced at higher sequencing depth. Sequence data from those colorectal cancer cases have been deposited at the European Genome-phenome Archive (EGA), which is hosted by the European Bioinformatics Institute and the Centre for Genomic Regulation, under accession no. EGAS00001003066. Further information about EGA can be found at https://ega-archive.org. Diploid SNVs and copy-number calls are available in the Supplementary Data. Data in Fig. 3f were from Lee et al.[24]. Data in Fig. 4 are available through the PCAWG consortium[25]. Whole-genome variant call data in Fig. 5, which were not available from the original publication, were provided upon email request by Korber et al.[28].

Code availability

MOBSTER is available as an R package at https://github.com/sottorivalab/mobster; future updates, as well as all vignettes and manuals, are maintained at https://caravagn.github.io/mobster. A repository with all Supplementary Data are available at https://github.com/sottorivalab/mobster_supp_data. Supplementary Data contain vignettes that show the analysis of single-sample and multiregion simulated tumors, the whole analysis of multiregion colorectal samples and single-sample lung cancers, and summary results from the PCAWG and GBM cohorts. Somatic SNVs and copy-number calls used for the analysis of multiregion colorectal samples are also available as Supplementary Data. The implementation of all other R packages that we have developed are available at https://caravagn.github.io/.

References

42. Fusco, D., Gralka, M., Kayser, J., Anderson, A. & Hallatschek, O. Excess of mutational jackpot events in expanding populations revealed by spatial Luria–Delbrück experiments. *Nat. Commun.* **7**, 12760 (2016).
43. Teh, Y. W. Dirichlet processes. in *Encyclopedia of Machine Learning* (eds Sammut, C. & Webb, G.) 280–287 (Springer, 2011).
44. Ghahramani, Z., Jordan, M. I. & Adams, R. P. Tree-structured stick breaking for hierarchical data. in *Advances in Neural Information Processing Systems* (eds Lafferty, J. D. et al.) 2319–2327 (Neural Information Processing Systems, 2010).
45. Ma, Z. & Leijon, A. Bayesian estimation of beta mixture models with variational inference. *IEEE Trans. Pattern Anal. Mach. Intell.* **33**, 2160–2173 (2011).
46. Clauset, A., Shalizi, C. R. & Newman, M. E. J. Power-law distributions in empirical data. *SIAM Rev.* **51**, 661–703 (2009).
47. Schröder, C. & Rahmann, S. A hybrid parameter estimation algorithm for beta mixtures and applications to methylation state classification. *Algorithms Mol. Biol.* **12**, 21 (2017).
48. Biernacki, C., Celeux, G. & Govaert, G. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**, 719–725 (2000).

Acknowledgements

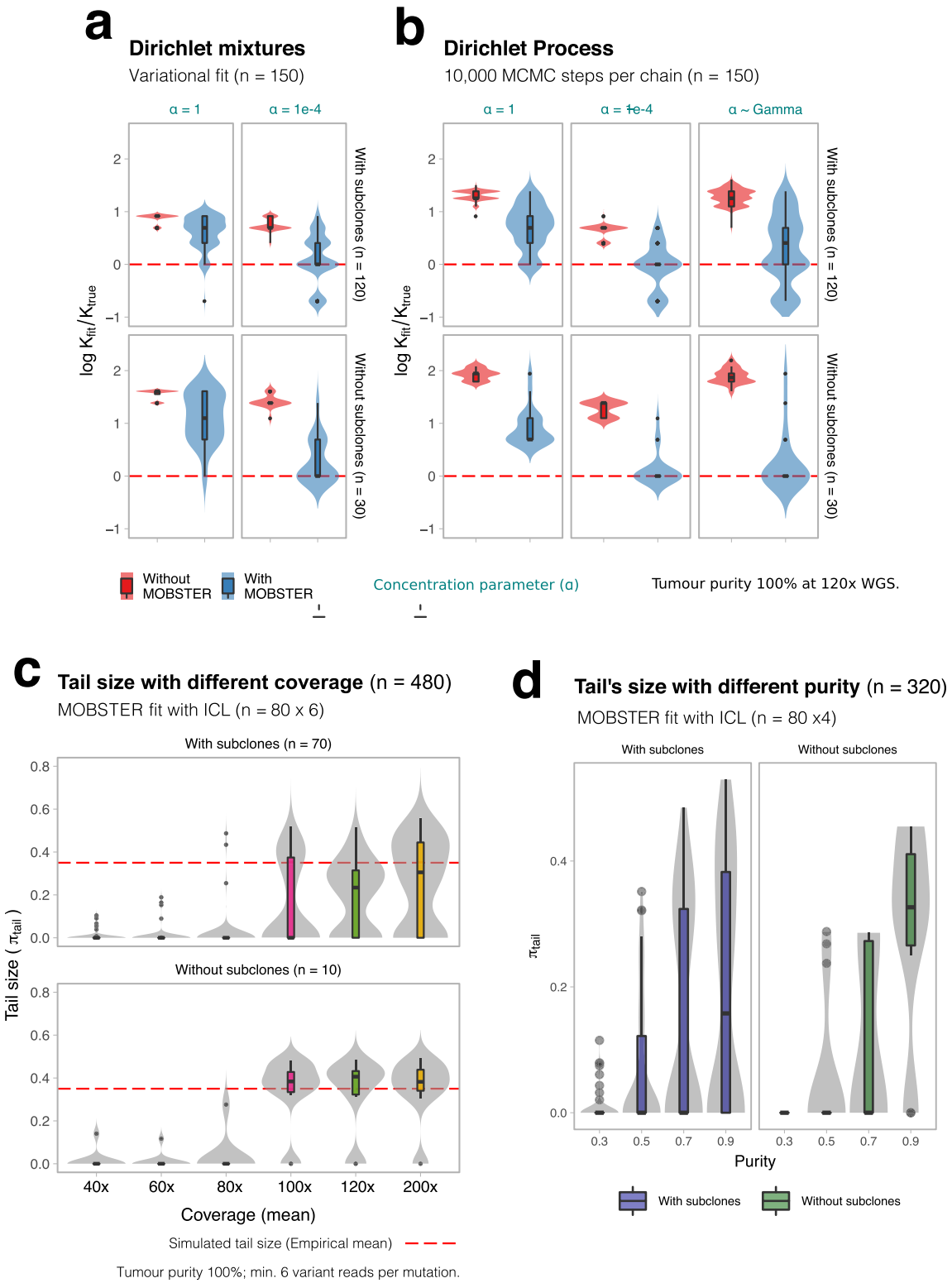
A.S. is supported by the Wellcome Trust (202778/B/16/Z) and Cancer Research UK (A22909). T.G. is supported by the Wellcome Trust (202778/Z/16/Z) and Cancer Research UK (A19771). We thank the Medical Research Council (MR/P000789/1) for funding A.S. and the National Institutes of Health (NCI U54 CA217376) for funding A.S and T.A.G. C.P.B. thanks the Wellcome Trust (209409/Z/17/Z) for funding. L.C. thanks Cancer Research UK (A24566) and Children with Cancer UK (17–235) for funding. This work was also supported by a Wellcome Trust award to the Centre for Evolution and Cancer (105104/Z/14/Z). L.Z. is supported by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska–Curie Research Fellowship scheme (846614). We thank N. Matthews and the Tumour Profiling Unit at the ICR for their support with next-generation sequencing. We thank V. Körber and T. Höfer for sharing their data and for fruitful discussion around the glioblastoma cohort.

Author contributions

G.C. conceived, designed and implemented the method. T.H. and K.C. developed the spatial tumor growth simulations. T.H. and M.W. generated the data for synthetic tests. G.C., T.H., M.W. and D.N. carried out and analyzed these tests. G.C., M.W. and L.Z. analyzed the data. W.C., G.D.C. and A.A. provided input and support with the analysis. G.S., C.B., T.A.G. and A.S. supervised the method design. L.C. contributed to study supervision. A.S. and T.A.G. conceived and supervised the study. All authors contributed to and approved the manuscript.

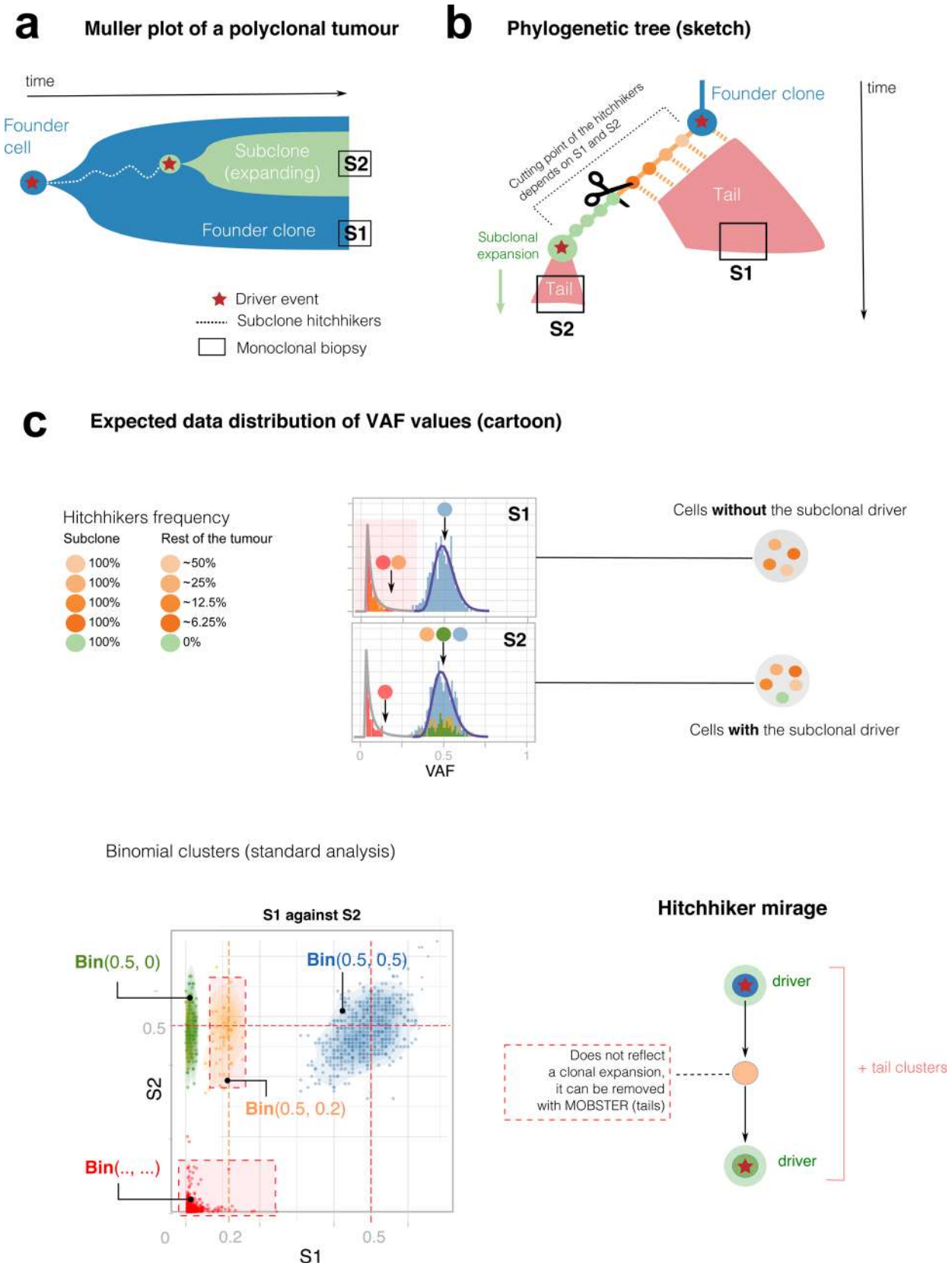
Competing interests

Authors declare no competing interests.



Extended Data Fig. 1 | See next page for caption.

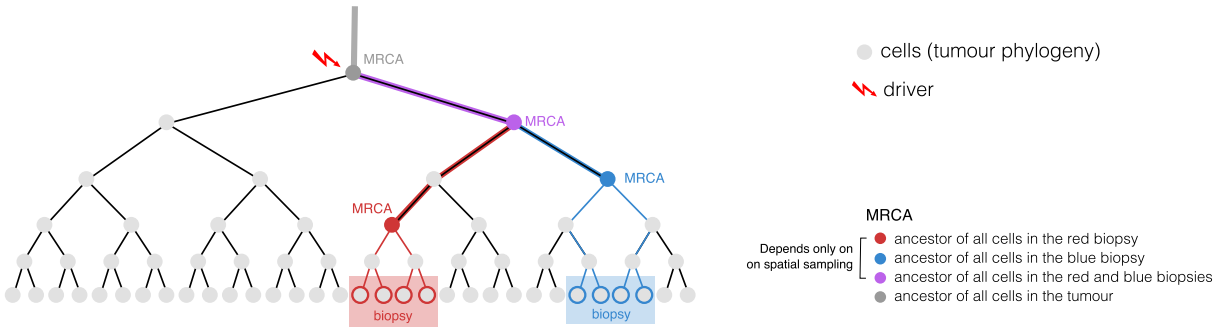
Extended Data Fig. 1 | Synthetic tests with MOBSTER. Example MOBSTER fit of synthetic single-sample tumors (details in Supplementary Note 1). All boxplots and violins show mean and inter quartile range (IQR), upper whisker is 3rd quartile +1.5 * IQR and lower whisker is 1st quartile - 1.5 * IQR. **a**,**b**, Subclonal reconstruction with MOBSTER, against standard methods (variational fit of a Dirichlet finite mixture, and a Markov Chain Monte Carlo sampling for a Dirichlet Process). These methodologies are at the basis of many approaches in the field. The test uses synthetic data from $n = 150$ simulated tumors ($n = 120$ with one subclone, and $n = 30$ without subclones), generated from a stochastic branching process. We report the logarithm of the ratio between the number of clones fit (k_{fit}) and the true number (k_{true}). Tests show different values of the concentration parameter α, which tunes the propensity to call clusters. Values (for example, $\alpha = 10^{-4}$) are point estimates, but we also test also a Dirichlet Process where α is learnt from the data using a Gamma prior. **c**, Proportion of mutations assigned to MOBSTER's tail changes with coverage, at fixed 100% tumor purity. We span coverage from 40x to 200x, using a subset of $n = 80$ tumors from the test in panels (**a**, **b**). The red dashed line is the median tail size across the test set (obtained from simulated tumor); tests suggest the coverage required to fit a tail. **d**, As for coverage, we tested with $n = 320$ tumors ($n = 80$ per configuration) the ability of detecting tails as a function of purity, fixing a coverage of 120x. The average tail size is reported (number of SNVs assigned to the tail in the fit).



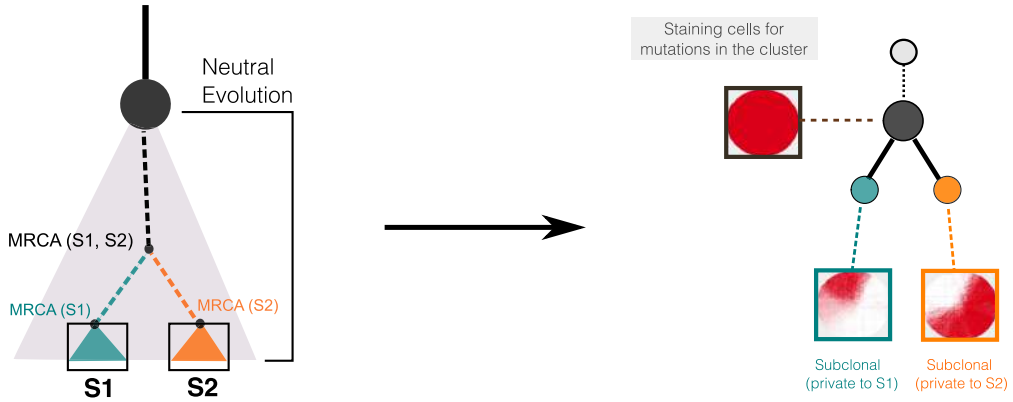
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | The hitchhiker mirage in multi-region sequencing data. a,**b**, Evolutionary history of a tumor with one subclone. After the first cancer cell gives rise to the tumor (blue founder clone), the population evolves neutrally accumulating passenger mutations (orange), until eventually a subclonal driver occurs triggering a new subclonal expansion (green, with its own tail). The subclonal driver, together with its passenger hitchhikers (orange) will rise in frequency with the subclonal expansion, forming a subclonal cluster in the VAF distribution. However, some early hitchhikers will also be present elsewhere in the tumor as part of the tail of the founder clone. In the example of perfect cell doubling, we expect mutations in the first doubling to be in 50% of the cells of the tumor, mutations in the second doubling to be in 25% etc. We take monoclonal biopsies S1 and S2, and find the founder clone (S1) and a subclonal sweep (S2). **c**, The hitchhiker mirage (Supplementary Note 2) is a confounder determined by passengers that hitchhike to the subclonal driver in S2, but diffuse neutrally in S1 (orange). This can be seen in the S1 vs S2 VAF scatter, where the orange mutations do not travel together in the two samples, because cells in S1 do not harbor the subclonal driver (while those in S2 do). The VAF scatter shows that orange hitchhikers can generate an extra cluster with Binomial parameters 0.5/0.2 for S1/S2, on top of the green clone with different parameters (S1/S=0.5/0). Moreover, extra clusters are generated by fitting tail mutations with a Binomial mixture, further inflating the true number of clones ($k=2$) and suggesting false clonal sweeps (from which the illusion of a non-existing clonal expansion). If we remove mutations assigned to a tail by MOBSTER we clean up the signal and retrieve the true clonal architecture.

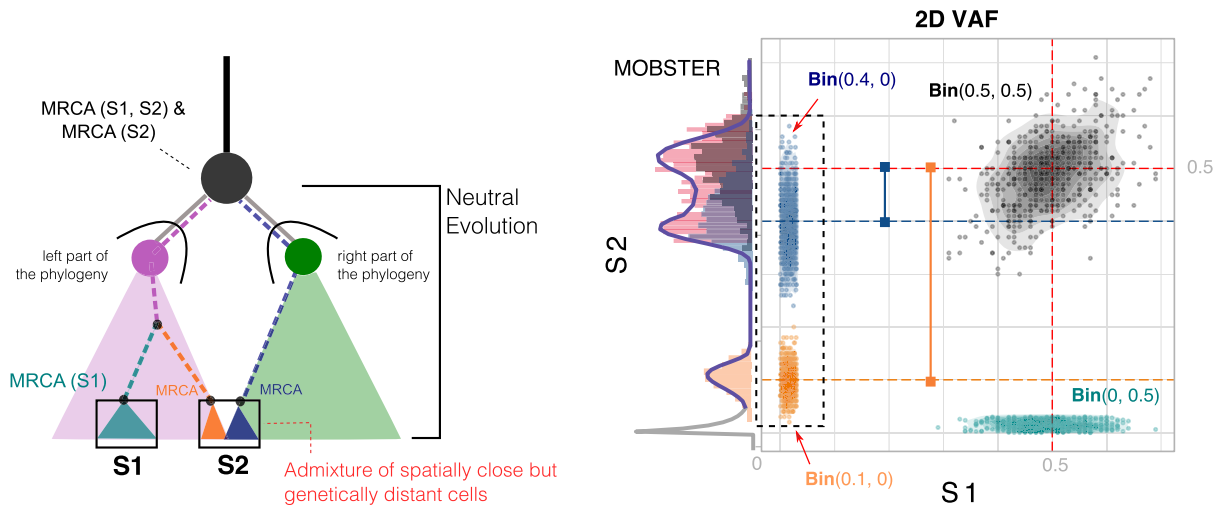
a **Example of Most Recent Common Ancestor (MRCA)**



b **MRCA effect and virtual staining matching the clone tree**

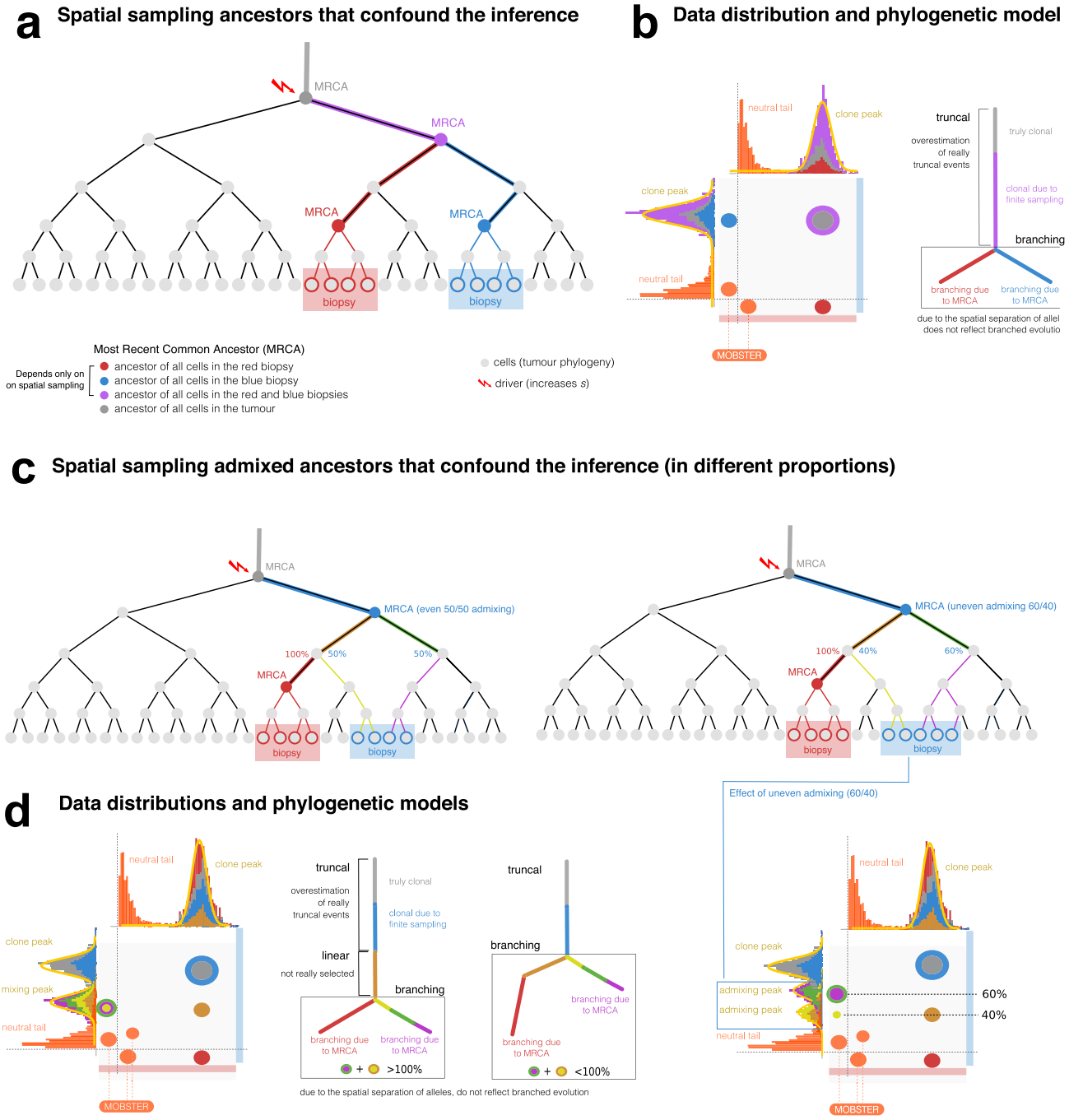


c **Admixing effect and expected data distribution of VAF values (cartoon)**

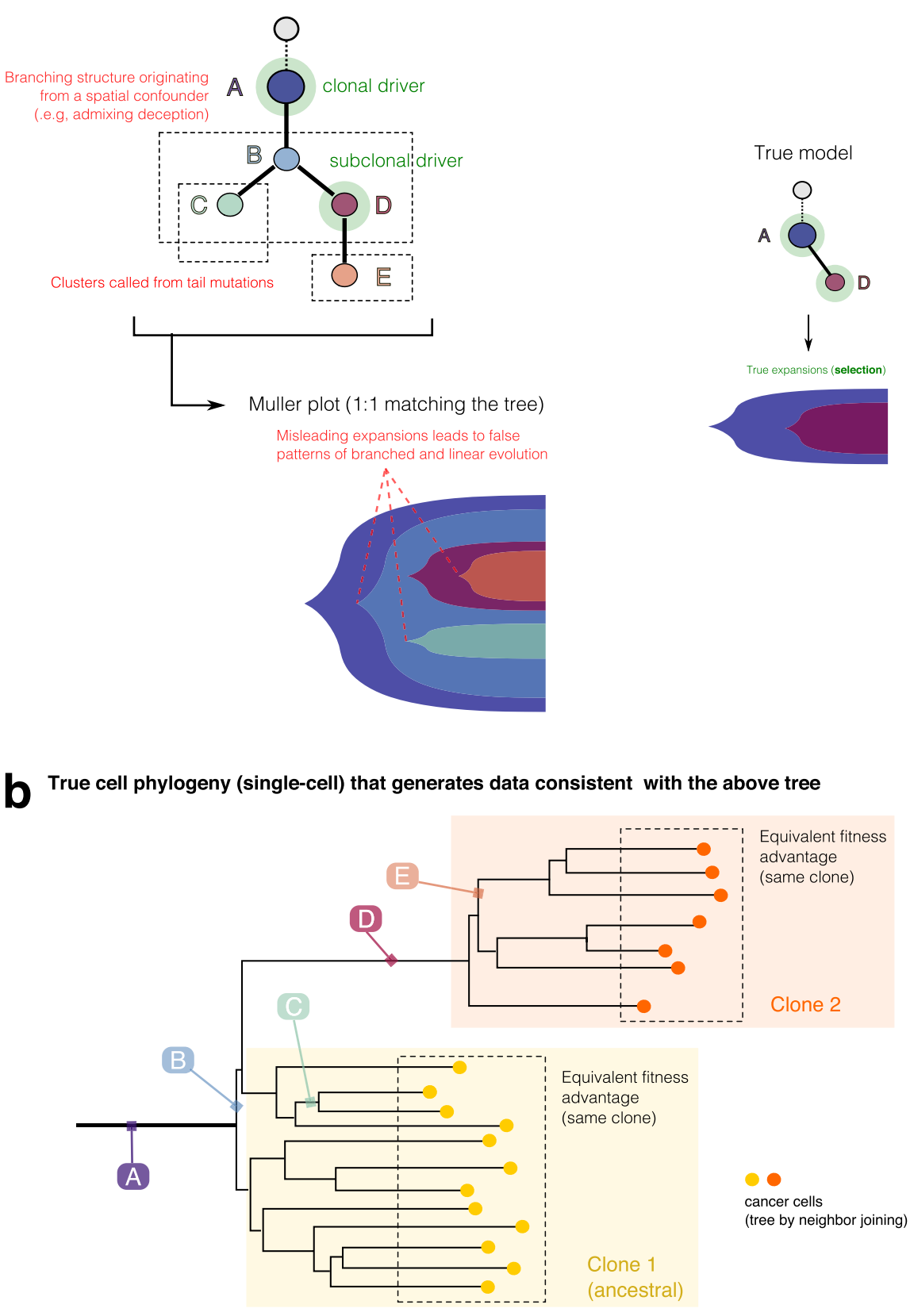


Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | The MRCA fallacy in multi-region sequencing data. a, Every cell always has an ancestor, and the cell starting the tumor is the Most Recent Common Ancestor (MRCA) of the whole tumor. We never sequence that cell, we sequence some of its progeny. We can travers the phylogeny of cell divisions backward, and determine the MRCA of all biopsy cells (red and blue), or the MRCA of all biopsies (purple). **b**, The ancestor effect (Supplementary Note 2) is the MRCA of cells from a spatially-localized biopsy, compared to other biopsies. Hence, mutations that are observed at high frequency in one biopsy are not necessarily due to selection. We simulate the growth of a 2D neutral tumor, and sample two biopsies (100% purity, S1 and S2). Both samples contain truncal mutations; each biopsy also contains private mutations (green and orange) that are clonal within the sample but are not due to selection. When we generate a virtual staining of all cells that harbor the mutations in a cluster, we see the separation between cells in S1 and S2, and the branched evolutionary structure in the clone tree that is not due to selection, but to spatial sampling (Extended Data Figure 4). **c**, The admixing deception stems from spatial tumor intermixing, with cells that are close in space, but genetically distant in the phylogeny. In this example, whereas S1 is a bulk of closely related cells, suffering only from the ancestor fallacy, S2 contains a mixture of cell lineages from distinct parts of the tree (here split in right and left). Intermixing is bond to happen since distant parts of a phylogeny must mix somewhere in space; again, in this example, no selection is at play. From these biopsies, we find truncal (black) and private mutations in S1(green, ancestor fallacy). In S2 we find a mixture of lineages (orange and blue) peaked like subclonal clusters (here we omit neutral tails for simplicity). The orange and blue clusters deviate by an offset that is determined by the level of admixing, which is unknown a priori (see the VAF of S2 in Extended Data Figure 4).



Extended Data Fig. 4 | Effects of the MRCA fallacy and the admixing deception in multi-region sequencing data. a, Phylogenetic tree of cellular divisions in a neutral expansion, that is, inside a clonal expansion triggered by a driver hitting the grey cell. The tree shows the sampling of 2 biopsies (red and blue), and the MRCAs. For example, the mutational load present in the red MRCA will characterize cells in the red biopsy. **b**, Data distribution and the associated phylogenetic tree show how our estimate of the true evolution of this tumor is confounded by spatial sampling. Mutations that accrue in the lineages from which the MRCAs originate, will create clusters in the data. The corresponding phylogenetic model will also show an inflated number of clonal events (purple MRCA), and branches that do not represent real selection-driven branched evolution (red and blue MRCAs). **c**, Example admixing deception in the blue biopsy, where two independent lineages are represented. Admixing can be even or uneven, depending on the proportion of lineages (left versus right) in the biopsy. Remark that no subclonal selection is at play in this example. **d**, If we sequence the above biopsies, we find truncal mutations (gray and blue), and a number of clusters that look like genuine subclones. The admixing effect is observed on the vertical of the blue biopsy. In the even case (50% each), the admixing generates one 50% peak for both independent lineages. According to the relation between the frequencies of the observed ancestors, we can also fit two different trees to data; notice that the branching structure presented in both of them is the result of the confounders and does not reflect actual branched evolution. In the uneven case (60% versus 40%), the two admixing peaks separate, originating 2 peaks hitting at the frequencies of 40% and 60%. This shows the pervasive effect of admixing, with up to 8 clusters in this simple scenarios.



Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Interpreting clone trees as clonal evolution models. Interpreting clone trees that contain spatial confounders as clonal evolution models can be difficult. We show an example consistent with the data shown in these Extended Data Figures 2–5. **a**, All the spatial confounders discussed in Supplementary Note 2 lead to additional nodes and branching structures in the estimated clone tree. These confounders need to be accounted for in a clonal deconvolution analysis, if we seek to identify waves of clonal expansions due to positive selection. The translation of clusters that originate from confounders, into clonal expansions due to selection is misleading, and the inferred clonal evolution is much more complex than the actual one. **b**, For clarification, a phylogenetic tree at the single cell level of the tumor, showing that clusters B, C, E and F are arbitrary ancestors identified by the specific spatial bias of the measurement.